Pre-Processing MRSes

Tore Bruland
Norwegian University of Science and Technology
Department of Computer and Information Science
torebrul@idi.ntnu.no

Abstract

We are in the process of creating a pipeline for our HPSG grammar for Norwegian (NorSource). NorSource uses the meaning representation Minimal Recursion Semantics (MRS). We present a step for validating an MRS and a step for pre-processing an MRS. The pre-processing step connects our MRS elements to a domain ontology and it can create additional states and roles. The pipeline can be reused by other grammars from the Delph-In network.

1 Introduction

NorSource[1] (Beermann and Hellan, 2004; Hellan and Beermann, 2005), a grammar for Norwegian, is a Head-Driven Phrase Structure Grammar (HPSG) (Sag et al., 2003), developed and maintained with the Linguistic Knowledge Builder (LKB) tool (Copestake, 2002), and originally based on the HPSG Grammar Matrix, which is a starter kit for developing HPSG grammars (Bender et al., 2002). An HPSG grammar can use Minimal Recursion Semantics (MRS) as meaning representation (Copestake et al., 2005). In order to speed up the parsing process (the unification algorithm), a HPSG grammar can be compiled and run (parsing) with the PET[2] tool (Callmeier, 2001). The Flop program in PET compiles the LKB grammar and the Cheap program runs it. An alternative to the PET system is the Answer Constraint Engine (ACE)[3] created by Woodley Packard. ACE can parse and generate using the compiled grammar.

Our goal is to create a pipeline for the NorSource grammar and use it to create small question-answer systems or dialogue systems. The first step in the pipeline is the parsing process with ACE. The next step is to select the most suitable MRS. We use Velldal's ranking model (Velldal, 2008). The model is based on relevant sentences from our system, treebanked with [tsdb++] (Oepen et al., 2002; Oepen and Flickinger, 1998). The selected MRS is checked with the Swiss Army Knife of Underspesification (Utool) (Koller and Thater, 2006b) and our own validating procedure. Only well-formed MRSes are used in our pipeline. We also use Utool to solve the MRS and to eliminate any logically equivalent readings. The next step is to pre-process the MRS (calculate event structure and generate roles), and the last step in our pipeline creates a First-Order Logic formula from the MRS (only the easy cases).

Our contribution is the validating step and the pre-processing step.

In the next section, we give a brief introduction to MRS. Then we present details from our validating procedure. Next, we solve an MRS and eliminate logically equivalent readings with Utool. We pre-process the selected MRS in section 6. At last, we look at a way to create a First-Order Logic formula from a solved MRS and we present a few challenges from our research.

[1] http://typecraft.org/tc2wiki/Norwegian_HPSG_grammar_NorSource

[2] http://pet.opendfki.de/

[3] http://moin.delph-in.net/AceTop

2 Minimal Recursion Semantics

The elements of an MRS can be defined by the structure mrs(T, I, R, C), where *T* is the top handle, *I* is the index, *R* is a bag of elementary predictions (EP), and *C* is a bag of constraints.

Every dog chases some white cat (1)

(1) can have the following MRS (created for demonstration purposes):

T h_0,

I e_1,

R { h_1:every(x_1,h_3,h_8), h_3:dog(x_1), h_7:white(x_2), h_7:cat(x_2), h_5:some(x_2,h_{10},h_9), h_4:chase(e_1,x_1,x_2) }

C $h_{10} =_q h_7$

An MRS can be in two states: unsolved or solved. An algorithm (LKB and Utool) brings an MRS from the unsolved state into one or more solved states. An unsolved MRS has holes that are not in the set of labels, and a solved MRS has holes that are from the set of labels. The set of labels in our example: $\{h_1$, h_3, h_4, h_5 and $h_7\}$. The set of holes: $\{h_3$, h_8, h_9 and $h_{10}\}$. A hole can be either open or closed. A hole is open when it isn't in the set of labels, and a hole is closed when the hole is in the set of labels. Hole h_3 is closed in our example.

3 Validate MRSes From NorSource

We want to search our MRSes for properties that can lead to problems. The Utool solvable function checks if an unsolved MRS can be transformed into one or more solved MRSes without violating the MRS definitions.[4] Our validating procedure contains a set of functions. We create variables or list of variables for each function that is positive. The functions are: *empty index*, *empty feature*, *empty reference*, *key conjunction*, and *argument EP conjunction*. An *empty index* exist when the index value refers to a variable that is not an EP's arg_0. An *empty feature* exist when a feature value refers to a variable that is not found in the EP's arguments. An *empty reference* exist when an argument refers to a variable that is not an EP's arg_0 and the variable is not in the set of feature values. A *key conjunction* exist when more than one EP in an MRS have the same arg_0 and they are not quantifiers. An *argument EP conjunction* exists when an argument contains a label that is an EP conjunction. In Table 1, the variable u_{12} is an empty feature. The variable u_{10} is a empty reference. The arg_0 of pred_2 and pred_3 form a key conjunction. The argument h_9 in arg_1 of pred_1 is an argument EP conjunction.

EP	Feature
h_3:pred$_1$(arg_0(e_1),arg_1(h_9))	e_1,feature$_1$,value$_1$
h_9:pred$_2$(arg_0(u_1),arg_1(x_1),arg_2(u_{10}))	u_{12},feature$_2$,value$_2$
h_9:pred$_3$(arg_0(u_1),arg_1(x_1),arg_2(x_2),arg_3(u_{15}))	u_{15},feature$_3$,value$_3$
h_2:pred$_4$(arg_0(x_1))	u_{16},feature$_4$,value$_4$
h_4:pred$_5$(arg_0(x_2))	

Table 1: Eps and Features

[4] Utool is stricter than the LKB software, see Fuchss et al. (2006).

4 Selecting An MRS

Before we create a ranking model, we analyze and compare the MRSes from our domain.[5] We use the variable-free solution Oepen and Lønning introduced (Oepen and Lønning, 2006). We compare parts from the syntax tree, the EPs, the features, and if the MRS is solvable or not. We also present results from our validation procedure. If we parse (2) with NorSource, it yields 9 MRSes. Parts of the EP information is presented in Table 2.

	1	2	3	4	5	6	7	8	9
legge_v ARG1 addressee-rel	x	x	x	x	x	x	x	x	x
legge_v ARG2 bok_n	x	x	x	x	x	x	x	x	x
legge_v ARGX på_p		x	x						
på_p ARG1 bok_n	x	x	x			x	x	x	x
på_p ARG1 legge_v				x	x				
på_p ARG2 bord-1_n	x	x	x	x	x	x	x	x	x

Table 2: Compare MRSes

Legg boken på bordet
Put the book on the table (2)

We use the LOGON software to treebank ([tsdb++]) and to create our ranking model (we have copied adjusted the scripts in folder lingo/redwoods). We parse with the ranking model and we select the first MRS.

5 Solving The MRS

Utool solves an MRS using a dominance graph and a chart (Niehren and Thater, 2003). The redundancy elimination algorithm (Koller and Thater, 2006a) takes a chart and a redundancy elimination file as input and returning the chart without the redundancy. We have created a redundancy elimination file according to (Koller and Thater, 2006b) for our quantifiers.

6 Pre-Processing The Selected MRS

In the pre-processing step we focus on the event structure and roles. By event structure we mean: sub events, aspectual and causal notions. Vendler grouped verbs into classes based on their temporal properties (Vendler, 1967). The verbs are classified according to duration and presence of a terminal point. A verb with a terminal point is called telic (the verb culminates). Vendler's classes are also known by other terms such as: eventualities, situations, lexical aspect or Aktionsart. The classes are: state, point, process, achievement, and accomplishment. The verb's connection to a class is not static, because a verb argument can move an event from one class into another. This phenomenon is called aspectual composition or coercion (Moens and Steedman, 1988).

A predicate string in an EP can contain the prefix "_", the suffix "_rel", a name, a part-of-speech type and a sense number. The name, the part-of-speech type and the sense number can be connected to a domain ontology definition. If we don't have the sense number, we can have a list of domain ontology definition candidates. A predicate string can also be a unique name like in "first_position_prominent".

Our goal with the pre-processing is:

- to connect the names in the predicate strings to a domain ontology
- to check if the predicate and the predicate arguments are valid according to the domain

[5] A demo for NorSource: `http://regdili.idi.ntnu.no:8080/comparemrs/compare`

- to create a common structure for a set of verbs
- implement an algorithm for roles and states

The main elements of our solution are a predicate tree, an algorithm, a domain ontology, and a set of object-oriented classes. The predicate tree is created from the predicates and their arguments.

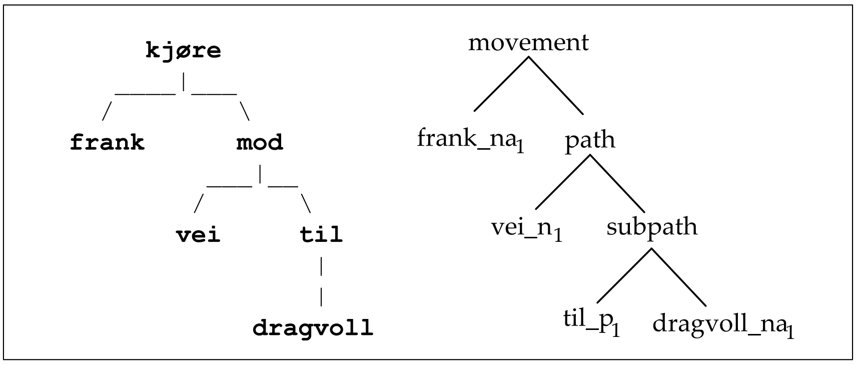


Figure 1: The Predicate Tree and the Movement Class

Frank kjører veien til Dragvoll (3)
Frank drives the road to Dragvoll

The predicate tree on the left in Figure 1 is created from (3). The algorithm searches the predicate tree and for each node in the tree it finds templates from the domain ontology. A template contains checks against the domain ontology, a return function and a return class. One of the templates used in our example is shown in Table 3. The variable X is replaced by vei_n$_1$ in our example. The final class for

key	nodes	check list	return	class

mod$_5$	node(n,X)
	node(subpath,til_p$_1$)
		isa(X,path_n$_1$)	new	path

Table 3: Template Example

(3) is shown on the right in Figure 1. We have defined three return functions: “new”, “call” and “fork”. “New” creates the return class from its arguments. “Call” is used when one node consumes another. For example in “VP PP”, the PP can be consumed by the VP. “Fork” is used when different classes are connected. For example in ”My uncle goes to town” we have a Family class and a Movement class that are connected through the variable for the uncle. The Movement class for (3) contains the following information:

movement
event(e_1:kjøre_v$_1$)
subject(x_2:frank_na$_1$)
path
object(x_9:vei_n$_1$)
end-point(x_1:dragvoll_na$_1$,t_2)

checks
has(dragvoll_na$_1$,location)
isa(vei_n$_1$,path_n$_1$)
templates
[tv$_2$, pp$_2$, mod$_5$]

The Movement class is a result from analyzing a number of different sentences about objects changing location. The Movement class is a process and it can contain a path, a departure, an arrival, a road (named path), a vehicle etc. A number of these objects are connected. For example the beginning of the process and the beginning of the path. The beginning of the path and the departure. We have implemented an algorithm for detecting the roles *work* and *cargo*. *Work* indicates an object that is using energy. *Cargo* indicates an object that is being transported.

The classes can be used as they are in a further reasoning process, or they can generate EPs and features that are inserted into the MRS. For example, the state *at_location(x_4,x_1)* can be inserted in an EP conjunction.

7 Logic Form

The solved MRSes are not yet formulas in First-Order Logic. We have to convert arguments that are in a Higher-Order Logic, insert the *not* operator, and use the quantifiers from First-Order Logic. An argument that has an Ep conjunction or has a handle needs to be rewritten. An argument with a handle is replaced with the *arg*$_0$ from the EP of the handle. We can give a reference and a predicate to the conjunction or we can find a candidate in the conjunction. We use the latter method. The predicate *_neg_adv_rel* is converted to the operator *not*. We have connected some of the NorSource quantifiers to the First-Order Logic quantifiers *exist* and *all*. Quantifiers like *some*, *few*, etc can be assigned to a group / type of their own. We place an operator between the arguments "RSTR" and "BODY" in our quantifiers:

$$\exists(y)[pred_1(y)\ \textbf{operator}_1\ \forall(x)[pred_2(x)\ \textbf{operator}_2\ pred_3(x,y)]]$$

We have defined **operator**$_1$ as $\wedge$ and **operator**$_2$ as $\rightarrow$. The formula from (3) is:

$$\exists(x_1)[\textit{na}(x_1,\textit{dragvoll})\wedge\exists(x_2)[\textit{na}(x_2,\textit{frank})\wedge\exists(x_9)[\textit{vei\_n}(x_9)\wedge\textit{kjøre\_v}(e_1,x_2,x_9)\wedge\textit{til\_p}(u_1,x_9,x_1)]]]$$

The meaning of the MRS quantifiers are preserved with the extra predicates: q_meaning(u100,x_1,def), q_meaning(u101,x_2,def) and q_meaning(u102,x_9,def). These predicates need to be inserted into the logic formula.

8 Challenges

The first challenge is to find a representative selection of sentences from the domain. The sentences are treebanked and used in our ranking model, and their selected MRSes are also used for creating domain ontology definitions. We use a virtual profile where we can add new annotated profiles.

We need to find a way to process our quantifiers that are not in First-Order Logic. At this early stage we can use off-the-shelf theorem provers and model finders as described in Blackburn's and Bos' reasoning framework for First-Order logic (Blackburn and Bos, 2005). We must classify different kinds of questions and commands, and we need to describe how to process them.

So far, we have connected types for each Ep together, but sometimes a more detailed structure is required in order to express meaning. A part of one structure can connect to a part of another. In the examples: "the cat sits in the car", "the cat sits on the car", and "the cat sits under the car", there are three different locations related to the car. We have a container in the car, an area on top of the car, and a space under the car. This extra information needs to be used together with the MRS. Sometimes more ambiguities are introduced and we need a ranking. For example in "Fred poured coffee on the thermos", the most probable is when the coffee ends up inside the thermos and the less probable version where the coffee is poured on the outside of the thermos. We also want to use previous situations and the discourse so far in our pre-processing step.

References

Beermann, D. and L. Hellan (2004). A treatment of directionals in two implemented hpsg grammars. In S. Müller (Ed.), *Proceedings of the HPSG04 Conference.* CSLI Publications.

Bender, E. M., D. Flickinger, and S. Oepen (2002). The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In

J. Carroll, N. Oostdijk, and R. Sutcliffe (Eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, pp. 8–14.

Blackburn, P. and J. Bos (2005). *Representation and Inference for Natural Language*. CSLI Publications.

Callmeier, U. (2001). *Efficient Parsing with Large-Scale Unification Grammars*. Master thesis, Universit¨at des Saarlandes.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI.

Copestake, A., D. Flickinger, C. Pollard, and I. A. Sag (2005). Minimal Recursion Semantics. An introduction. *Journal of Research on Language and Computation 3*(4), 281 – 332.

Fuchss, R., A. Koller, J. Niehren, and S. Thater (2006). Minimal recursion semantics as dominance contraints: Translation, evaluation and analysis. In *Proceedings of the 42nd ACL*. Association for Computational Linguistics.

Hellan, L. and D. Beermann (2005). Classification of prepositional senses for deep grammar applications. In V. Kordoni and A. Villavicencio (Eds.), *Proceedings of the 2nd ACL-Sigsem Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Colchester, United Kingdom.

Koller, A. and S. Thater (2006a). An improved redundancy elimination algorithm for underspecified representations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 409–416. Association for Computational Linguistics.

Koller, A. and S. Thater (2006b). *Utool: The Swiss Army Knife of Underspesification*. Saarland University, Saarbrücken.

Moens, M. and M. Steedman (1988). Temporal ontology and temporal reference. *Computational Linguistics 14*(2).

Niehren, J. and S. Thater (2003). Bridging the gap between underspesification formalisms: Minimal recursion semantics as dominance constraints. In *41st Meeting of the Association of Computational Lingustics*, pp. 367–374.

Oepen, S. and D. Flickinger (1998). Towards systematic grammar profiling. test suite technology ten years after. *Journal of Computer Speech and Language 12*(4), 411–436.

Oepen, S. and J. Lønning (2006). Discriminant-based mrs banking. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Oepen, S., K. Toutanova, S. Shieber, C. Manning, D. Flickinger, and T. Brants (2002). The lingo redwoods treebank: Motivation and preliminary applications. In *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, pp. 1–5. Association for Computational Linguistics.

Sag, I. A., T. Wasow, and E. M. Bender (2003). *Syntactic Theory: a formal introduction* (2. ed.). CSLI Publications.

Velldal, E. (2008). *Empirical realization ranking*. Ph. D. thesis, The University of Oslo.

Vendler, Z. (1967). *Linguistics in Philosophy*. Cornell Univerity Press.

The semantic annotation of quantification

Harry Bunt
TiCC, Tilburg Center for Cognition and Communication
Tilburg University, The Netherlands
harry.bunt@uvt.nl

Abstract

This paper presents an approach to the annotation of quantification, developed in the context of an ISO project aiming to define standards for semantic annotation. The approach is illustrated for a range of quantification phenomena, including cumulative, collective, and group quantification.

1 Introduction

In 2012, two ISO standards for semantic annotation were established, one for time and events (ISO 24617-1), and one for dialogue acts (ISO 24617-2); others are under development for semantic roles, spatial information, and discourse relations. Quantification turns up as a problem in nearly all of these efforts. ISO 24617-1 has some provisions for dealing with quantification (see Pustejovsky et al., 2010), but these are too limited and do not always give correct results (Bunt and Pustejovsky, 2010).

The annotation of quantification faces three main issues:

1. Which set of semantic features, expressed most likely as XML attributes and values, adequately characterize a wide range of forms of quantification;
2. Is it possible to define a formal semantics of these expressions;
3. How can the relevant features (attributes and values) be defined as components of the semantic annotations in a way that respects compositional semantics?

Bunt (2005) proposed a way to representing quantifications in terms of feature structures, however, in this proposal the properties of a quantification are all expressed as properties of an event, which is inconvenient for annotators and is hard to combine with a compositional NP semantics. This paper presents an approach where quantification features are distributed over annotation structure components in a way that corresponds to their linguistic expression in syntactic structures, and shows the semantic adequacy of the proposal by a compositional translation into discourse representation structures.

2 Aspects of quantification

Quantification occurs when a predicate is applied to sets of arguments. Questions then arise concerning the precise way that the predicate is applied to members of these sets. As an example, consider the sentence (1a); some of the questions that may be asked (and answered) are (1b) - (1f):

(1) a. *Although a threat had been made before, three men rather unexpectedly moved both pianos.*
b. How many men were involved? (Answer: *Three.*)
c. How many pianos were involved? (Answer: *Two.*)
d. Did the same men move both pianos? (Answer: *Yes.*)
e. Did the men act collectively or individually? (Answer: *Collectively, probably.*)
f. Were the pianos acted upon collectively or individually? (Answer: *Individually, probably.*)

Given the *restriction* part of a quantification, that specifies a domain from which elements can be taken that participate in certain events individually, in groups, or collectively, the *distribution* is a function that computes the set of those entities that participate in the events as agents, as themes, as instruments, etc. We call such a set of participants a *predication domain*, and the domain, defined by the restriction, the *reference domain*. In the case of individual distribution, the two are the same.

A common function of proportional quantifier words like *all, some,* and *most*, and of absolute quantifiers like *three, more than five, 2 litres of* is to specify the fraction of the reference domain that is involved in the events under consideration. Numerical and amount quantifiers may also be used to indicate the size of a reference domain, like *twelve* in *The twelve students in this room all speak two languages.*

Proportional and absolute quantifiers can also be used to indicate the number/amount of a predication domain per element of another predication domain, like *five* in *Each of the dogs ate five sausages*.

Some of the most important aspects of quantification, distinguished in (Bunt, 1985) are:

(2)
1. the quantifier's restriction, describing the reference domain of the quantification;
2. the distribution, defining the predication domain;
3. size of the reference domain;
4. involvement of the reference domain (in absolute or relative terms);
5. relative scoping of the quantifications associated with argument NPs;
6. scoping of NP-quantifications relative to quantified events;
7. size of groups of elements from a reference domain;
8. number of elements of a reference domain involved per element of a predication domain.

3 Events, participants, and quantification annotation

In a davidsonian approach to meaning, we may view the combination of a verb with its arguments as introducing a set of events ('eventualities', more generally) of the type indicated by the verb, and with a number of properties concerning the way in which the participants of these events are involved. Applying this view to sentence (1a), we obtain a descriptions in terms of a set of move-events, a set of men, participating in these events as agents, and a set of pianos that participate as themes. This is exactly what is expressed in an annotation of semantic roles according to the ISO standard 24617-4 under development (see Bunt & Palmer, 2013):

```
      [v1:] <event xml:id="e1" target="#m2" eventType="move"/>
      [p1:] <participant xml:id="x1" target="#m1" entityType="man"/>
(3)   [p2:] <participant xml:id="x2" target="#m3" entityType="piano"/>
      [L1:] <srLink event="#e1" participant="#x1" semRole="agent"/>
      [L2:] <srLink event="#e1" participant="#x2" semRole="theme"/>
```

Since quantification occurs when two or more sets of arguments are related by a predicate, we can view the quantifications in (1a) as due to the Agent and Theme predicate relating sets of events and participants. This information can thus be represented as properties of the semantic role links (`<srLink>`), adding features to these links as follows, where the feature `@signature="set"` expresses that the sentence is about sets of events and participants :

```
      [v1:] <event xml:id="e1" target="#m2" eventType="move" signature="set"/>
      [p1:] <participant xml:id="x1" target="#m1" entityType="man”
             signature ="set" involvement="3"/>
      [p2:] <participant xml:id="x2" target="#m3" entityType="piano"
(4)          signature ="set" definiteness="def" involvement="2"/>
      [L1:] <srLink event="#e1" participant="#x1" semRole="agent"/>
             distribution="collective"/>
      [L2:] <srLink event="#e1" participant="#x2" semRole="theme"
             distribution="individual"/>
```

The correctness and usefulness of annotating quantification this way depends on how well it deals with the three issues mentioned in Section 1: expressive adequacy of attributes and values; their semantic adequacy; and compatibility with compositional phrase semantics. These issues are addressed next.

4 Representational and semantic adequacy

Of the information types listed in (2), those numbered 1, 2, 4, and 8 can be represented by the attributes and values shown in (4). For types 3, 5, 6 and 7, the attribute `@cardinality` is defined for `<event>` (for *"say twice"* etc.) and `<participant>` elements; the attribute `@outScoping` allows the expression of relative scope restrictions; and the values of the `@groupCard` attribute can be used to indicate group sizes in group quantifications. This provides the expressive power to represent a wide range of quantification phenomena.

The issue of compatibility with compositional NP semantics arises because we propose to represent some of the properties of a quantification as parts of semantic role links, where traditionally the semantic representation of quantification is considered to be part of NP semantics. This is in particular the case with distribution, as in (4). Having `@distribution` as an attribute of `<participant>` elements would run into problems for a sentence such as *The men had a beer before they moved the piano*, where the subject NP should be interpreted individually for the drinking, but collectively for the lifting.

The semantic adequacy of the proposed annotation format can be shown by defining a systematic translation of annotations into DRSs, following Bunt (2011; 2013). XML elements describing sets of events or participants, like those in (4), are translated to a DRS which introduces a higher-order (i.e. set-valued) discourse marker,[1] and conditions translating the other features. DRSs interpreting linking elements introduce discourse markers for the linked sets of elements, and conditions that further characterize the link, e.g.:

[v1:] $\rightsquigarrow \langle\{E_1\},\{\langle\langle\{e\},\{e \in E_1\}\rangle \Rightarrow \langle\{\},\{move(e)\}\rangle\rangle\}\rangle$, or in box form:

$$\begin{array}{|l|}\hline E1 \\ \hline \begin{array}{|l|}\hline e \\ \hline e \in E \\ \hline \end{array} \Rightarrow \begin{array}{|c|}\hline \\ \hline \text{MOVE}(e) \\ \hline \end{array} \\ \hline \end{array}$$

$$[\text{p1}]\!: \rightsquigarrow \begin{array}{|l|}\hline X \\ \hline card(X)\text{=}3 \\ \begin{array}{|l|}\hline x \\ \hline x \in X \\ \hline \end{array} \Rightarrow \begin{array}{|c|}\hline \\ \hline \text{MAN}(x) \\ \hline \end{array} \\ \hline \end{array} \qquad [\text{L1:}] \rightsquigarrow \begin{array}{|l|}\hline E, X \\ \hline \begin{array}{|l|}\hline e \\ \hline e \in E \\ \hline \end{array} \Rightarrow \begin{array}{|c|}\hline \\ \hline \text{AGENT}(e,X) \\ \hline \end{array} \\ \hline \end{array}$$

The merge of these DRSs plus those translating [p2] and [L2] yields the satisfactory result:[2]

$$[\text{p1} \cup \text{L1}] \cup [\text{p2} \cup \text{L2}] \cup \text{v1} = \begin{array}{|l|}\hline E, X, Y \\ \hline card(X)\text{=}3 \quad card(Y)\text{=}2 \\ \begin{array}{|l|}\hline e \\ \hline e \in E \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline \text{MOVE}(e) \\ \text{AGENT}(e,X) \\ \hline \end{array} \quad \begin{array}{|l|}\hline y \\ \hline y \in Y \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline \text{e} \\ \hline e \in E \\ \text{PIANO}_{cr}(y) \\ \text{THEME}(e,y) \\ \hline \end{array} \\ \begin{array}{|l|}\hline x \\ \hline x \in X \\ \hline \end{array} \Rightarrow \begin{array}{|c|}\hline \\ \hline \text{MAN}(x) \\ \hline \end{array} \\ \hline \end{array}$$

Note that in this example both NPs outscope the verb; their relative scoping is irrelevant since the group of three men acted collectively, as a single entity. While verbs very often have narrow scope relative to argument NPs, this is not always the case, as (5) illustrates. Using the attributes and values introduced so far, the wide-scope interpretation of *"die"* is easily annotated:[3]

(5) a. *Everybody will die.*
b. Annotation for wide-scope 'die':

[1] The presentation is simplified here; see Bunt (2013) for the use of pairs $\langle m,x\rangle$ of markables and discourse markers, where the markables make sure that only the intended marker variables are unified upon DRS-merging.

[2] The subscript 'cr' (for 'contextually relevant') indicates the interpretation of the definiteness of the NP *both pianos*.

[3] We assume here an approach to semantic roles which allows an event to have more than one theme, such as the LIRICS annotation scheme (Petukhova & Bunt, 2007).

```
[v1:] <event xml:id="e1" target="#m1" eventType="die" signature="set"
      cardinality="1"/>
[p1:] <participant xml:id="x1" target="#m2" entityType="person"
      signature="set" involvement="all"/>
[L1:] <srLink event="#e1" participant="x1" semRole="theme"
      outScoping="#e1 #x1"/>
```

c. Semantics:

$$\begin{array}{|l|}\hline E,X \\ \hline card(E)=1 \quad \begin{array}{|l|}\hline x \\ \hline e\in E \\ \hline\end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline \text{DIE}(e) \\ \hline\end{array} \\ \\ \begin{array}{|l|}\hline x \\ \hline \text{PERSON}(x) \\ \hline\end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline \text{THEME}(e,x) \\ \hline\end{array} \\ \hline\end{array}$$

Cases of quantified NPs with unequal scope are readily annotated in the format outlined here, but in order to support these annotations by a well-defined semantics a new kind of merge operation on DRSs is neede, the *scoped merge*, which is defined as follows:

(6)
$$\begin{array}{|l|}\hline \delta \\ \hline \varepsilon \\ \begin{array}{|c|}\hline \alpha_1 \\ \hline \alpha_2 \\ \hline\end{array} \Rightarrow \begin{array}{|c|}\hline \beta_1 \\ \hline \beta_2 \\ \hline\end{array} \\ \hline\end{array} \oplus \gamma \;=\; \begin{array}{|l|}\hline \delta \\ \hline \varepsilon \\ \begin{array}{|c|}\hline \alpha_1 \\ \hline \alpha_2 \\ \hline\end{array} \Rightarrow (\begin{array}{|c|}\hline \beta_1 \\ \hline \beta_2 \\ \hline\end{array} \cup \gamma) \\ \hline\end{array}$$

For instance, applied to the classical scoping example (7a), we obtain the semantics (7c) of the annotation (7b):

(7) a. *All the students read two papers*

b. Annotation for wide-scope *two papers*:

```
[v1:] <event xml:id="e1" target="#m1" eventType="read" signature="set"/>
[p1:] <participant xml:id="x1" target="#m2" entityType="student"
      definiteness="def" signature ="set" involvement="all"/>
[p2:] <participant xml:id="x2" target="#m3" entityType="paper"
      signature ="set" cardinality="2" involvement="all"/>
[L1:] <srLink event="#e1" participant="x1" semRole="agent"
      distribution="individual" outScoping="#x1 #e1"/>
[L2:] <srLink event="#e1" participant="x2" semRole="theme"
      distribution="individual" outScoping="#x2 #x1"/>
```

c. Semantics:

$$[\text{p2} \cup \text{L2}] \oplus [\text{p1} \cup \text{L1}\cup v\text{1}] = \begin{array}{|l|}\hline Y \\ \hline card(Y)=2 \quad \begin{array}{|l|}\hline y \\ \hline y\in Y \\ \hline\end{array} \Rightarrow \begin{array}{|l|}\hline E,X \\ \hline \text{PAPER}(y) \quad \begin{array}{|l|}\hline x \\ \hline x\in X \\ \hline\end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline \text{STUDENT}_{cr}(x) \\ \hline\end{array} \\ \\ \begin{array}{|l|}\hline e \\ \hline e\in E \\ \hline\end{array} \Rightarrow \begin{array}{|ll|}\hline x & \\ \hline \text{READ}(e) & \text{AGENT}(e,x) \\ x\in X & \text{THEME}(e,y) \\ \hline\end{array} \\ \\ \begin{array}{|l|}\hline x \\ \hline x\in X \\ \hline\end{array} \Rightarrow \begin{array}{|ll|}\hline e & \\ \hline e\in E & \text{AGENT}(e,x) \\ \hline\end{array} \\ \hline\end{array} \\ \hline\end{array}$$

Besides scoped quantifications, also unscoped, partially scoped, and equally scoped quantifications have to be considered. Partially and unscoped scoped cases, where there is no or incomplete information about relative scoping, are easily annotated by not specifying values for `@outScoping` attributes, and interpreted with underspecified DRSs (see Reyle 1993; 1994). Equally scoped quantifications, as occurring in *cumulative quantification* (Scha, 1981) and in group quantification (Bunt, 1985), can be annotated using the attribute `@eqScope`, as shown in (8) and (9). The semantics is obtained simply by using the ordinary rather than the scoped merge of the sub-DRSs.

(8) a. *Three breweries supplied fifteen inns.*

b. Annotation of cumulative reading (*In total 3 breweries supplied in total 15 inns*):

```
[v1:] <event xml:id="e1" target="#m2" eventType="supply" signature="set"/>
[p1:] <participant xml:id="x1" target="#m1" entityType="brewery"
        signature ="set" involvement="3"/>
[p2:] <participant xml:id="x2" target="#m3" entityType="inn"
        signature ="set" involvement="15"/>
[L1:] <srLink event="#e1" participant="#x1" semRole="agent"
        distribution="individual" eqScope="#x1 #x2"/>
[L2:] <srLink event="#e1" participant="x2" semRole="theme"
        distribution="individual" eqScope="#x2 #x1"/>
```

c. Semantics:

$$(\text{p1} \cup \text{L1}) \cup (\text{p2} \cup \text{L2}) \cup \text{v1} = \begin{array}{|l|}\hline E, X, Y \\ \hline \text{card}(X) = 3 \quad \text{card}(Y) = 5 \\ \begin{array}{|c|}\hline x \\ \hline x \in X \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline e \\ \hline \text{BREWERY}(x) \\ e \in E \\ \text{AGENT}(e,x) \\ \hline \end{array} \qquad \begin{array}{|c|}\hline y \\ \hline y \in Y \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline e \\ \hline \text{INN}(y) \\ e \in E \\ \text{THEME}(e,y) \\ \hline \end{array} \\ \\ \begin{array}{|c|}\hline e \\ \hline e \in E \\ \hline \end{array} \Rightarrow \begin{array}{|ll|}\hline x, y & \\ \hline \text{SUPPLY}(e) & \\ x \in X & y \in Y \\ \text{AGENT}(e,x) & \text{THEME}(e,y) \\ \hline \end{array} \\ \hline \end{array}$$

Group quantification, as in *Three boys played soccer against five girls* on the reading where teams of 3 boys played against teams of 5 girls, can be annotated by using, besides the `@eqScope` attribute, the `@groupCard` attribute. The semantics is obtained by interpreting the participant annotations as DRSs introducing discourse markers for sets of sets of 3 boys and 5 girls.

```
      [e1:] <event xml:id="e1" target="#m2" eventType="play" signature="set"/>
      [p1:] <participant xml:id="x1" target="#m1" entityType="boy"
              signature ="set" groupCard="3"/>
      [p2:] <participant xml:id="x2" target="#m3" entityType="girl"
(9)           signature ="set" distribution="group" groupCard="5"/>
      [L1:] <srLink event="#e1" participant="x1" semRole="agent"
              distribution="group" eqScope="#x1 #x2"/>
      [L2:] <srLink event="#e1" participant="x2" semRole="co-agent"
              distribution="group" eqScope="#x2 #x1"/>
```

5 Concluding Remarks

We have described the essentials of a way of annotating quantification phenomena where the features that characterize different forms of quantification and their properties are distributed over components of annotation structures in a way that corresponds to their linguistic expression (e.g., involvement and cardinality are features of `<participant>` components, corresponding to their expression in NPs).

Taking a davidsonian approach, we have introduced attributes and values for events, event participants, and the semantic roles of participants in events. However, the view of quantification which underlies this is more general; when predicate and argument structures are used, rather than events with participants and semantic roles, that is easily accommodated, by introducing `<predicate>`. `<argument>`, and `<argLink>` elements, and e.g. attaching distribution and scoping features to the letter, as in:

```
(10) <argLink pred="#P1" arg="#x1" argNum="arg1" distr="collective"/>
```

Quantification over events, time and place can be annotated in a similar way. As suggested by Lee & Bunt (2012), Temporal quantification can be annotated by adding features to the `<event>` and `<timex3>` elements defined in ISO 24617-2 and to the `<tLink>` that represents the semantic relation between a set of events and a set of temporal entities (as in *Most of the professors teach every Monday*.

Clearly, something similar can be done for the annotation of quantifications over events and space, as occurring in *Policemen can be found at every streetcorner*, adding the same features to the elements defined in the ISO-Space language under development (see Pustejovsky et al., 2012).

References

Bunt, H. (1985). *Mass Terms and Model-Theoretic Semantics.* Cambridge University Press.

Bunt, H. (2005). Quantification and Modification Represented as Feature Structures. In *Proceedings 6th International Workshop on Computational Semantics (IWCS-6)*, Tilburg, Netherlands, pp. 54–65.

Bunt, H. (2011). Abstract syntax and semantics in semantic annotation, applied to time and events. In E. Lee and A. Yoon (Eds.), *Recent Trends in Language and Knowledge Processing*. Seoul: Hankukmunhwasa.

Bunt, H. (2013). A Methodology for Designing Semantic Annotations. *Language Resources and Evaluation (forthc.).*

Bunt, H. and M. Palmer (2013). Conceptual and Representational Choices in Defining an ISO Standard for Semantic Role Annotation. In *Proceedings of the Ninth Joint ACL-ISO Workshop on Interoperable Semantic Annotation ISA-9*, Potsdam.

Bunt, H. and J. Pustejovsky (2010). Annotation of temporal and event quantification. In *Proceedings of the Fifth Joint ACL-ISO Workshop on Interoperable Semantic Annotation ISA-5*, Hong Kong, pp. 15–22.

ISO24617-1:2012 (2012). *Language Resource Management - Semantic Annotation Framework, Part 1: Time and Events (SemAF/Time).* ISO Standard 24617-1, March 2012, Geneva. March 2012.

ISO24617-2:2012 (2012). *Language Resource Management - Semantic Annotation Framework, Part 2: Dialogue Acts (SemAF/Dialogue acts).* ISO Standard 24617-2, September 2012, ISO, Geneva.

Lee, K. and H. Bunt (2012). Counting time and events. In *Proceedings of the Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation ISA-8*, Pisa, pp. 34–41.

Petukhova, V. and H. Bunt (2008, May). LIRICS Semantic Role Annotation: Design and Evaluation of a Set of Data Categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Genova, Italy.

Pustejovsky, J., K. Lee, H. Bunt, and L. Romary (2010). ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Sevenfth International Conference on Language Resources and Evaluation (LREC 2010), Malta*, Paris, pp. 394–397. ELDA.

Pustejovsky, J., J. Moszkowics, and M. Verhagen (2012). The Current Status of ISO-Space. In *Proceedings of Joint ISA-7, SRSL-3 and I@MRT Workshop at LREC 2012), Istanbul*, Paris. ELDA.

Reyle, U. (1993). Dealing with Ambiguity by Underspecification. *Journal of Semantics 10(2)*, 123–179.

Reyle, U. (1994). Monotonic Disambiguation and Plural Pronoun Resolution. In H. Kamp (Ed.), *Ellipsis, Tense, and Questions. Esprit project 6852 DYANA-2, Deliverable R2.2.B.* Stuttgart: IMS.

Scha, R. (1981). Distributive, Collective and Cumulative Quantification. In J. Groenendijk and M. Stokhof (Eds.), *Formal Methods in the Study of Language*. Amsterdam: Mathematical Center.

Scope Disambiguation as a Tagging Task

Kilian Evang
CLCG, Rijksuniversiteit Groningen
k.evang@rug.nl

Johan Bos
CLCG, Rijksuniversiteit Groningen
johan.bos@rug.nl

Abstract

In this paper we present a pragmatic account of scope alternation involving universal quantifiers in a lexicalist framework based on CCG and DRT. This account can derive the desired reading for 96% of all cases of scope interaction involving universal quantification mediated by prepositions in a real corpus. We show how this account allows for recasting scope resolution as a simple token classification task, providing a simpler handle for statistical approaches to scope resolution than previous accounts.

1 Introduction

The correct handling of scope-bearing operators such as quantifiers, negation and intensional verbs is crucial for constructing logical form from text that capture the intended meaning. A lot of work has been done to describe the scope alternation behavior especially of quantifiers and to construct underspecified meaning representations from which all (theoretically) possible readings of a sentence containing them can be enumerated (Reyle, 1993; Bos, 1996; Copestake et al., 2005; Egg et al., 2001). The problem of finding the preferred reading in each concrete case has also been addressed, using machine learning (Higgins and Sadock, 2003; Andrew and MacCartney, 2004; Srinivasan and Yates, 2009; Manshadi and Allen, 2011).

Despite these efforts, existing wide-coverage semantic parsers do not typically resolve scope. Either the semantic representations they output are shallow and do not use logical quantifiers, which is sufficient for many applications but not for model-theoretic reasoning. Or they leave the output underspecified or always deterministically pick the same scoping behavior for a given quantifier.

How to make a semantic parser aware of scope alternation? We argue that an additional layer of tags is a good way. A semantic parser can use these tags to guide its decisions—just like it might use a layer of word sense tags and a layer of named-entity tags, and just like a syntactic parser uses part-of-speech tags. Scope tags can guide the construction of logical forms from the start and thereby avoid the additional computational and representational complexity of underspecified representations. At the same time, this approach avoids the need to change the syntactic parsing model.

In this paper we show how we applied this technique to the semantic parsing system of Curran et al. (2007). It consists of C&C, a statistical syntactic parser whose output includes predicate-argument structures, and Boxer, a semantic construction component which builds interpretable semantic representations. The resulting enhanced system is being used in constructing the Groningen Meaning Bank (GMB), a large corpus of English text annotated with logical form (Basile et al., 2012).

2 Formal Background

Boxer's lexical-semantic framework (Bos, 2009) uses the output of the C&C parser, which consists of derivations in Combinatory Categorial Grammar (CCG; Steedman, 2001). In these derivations, words and smaller constituents combine into larger constituents using combinatory rules such as forward application, which combines a *functor* category $X/Y : f$ with an *argument* category $Y : a$ to yield a new constituent with category $X : f@a$ (1-place function f applied to a). Left of the colon are syntactic

$$\lambda p.(\begin{array}{|c|}\hline x\\\hline \text{member}(x)\\\hline\end{array};(p@x)) \qquad \lambda p.\begin{array}{|l|}\hline \\\hline \begin{array}{|c|}\hline x\\\hline \text{member}(x)\\\hline\end{array}\Rightarrow (p@x)\\\hline\end{array}$$

Figure 1: Generalized quantifiers denoted by the NPs *a member* and *every member*

$$\lambda io.\lambda do.\lambda su.\lambda mp.(su@\lambda x.(io@\lambda y.(do@\lambda z.(\begin{array}{|c|}\hline e\\\hline \text{give}(e)\\ \text{Agent}(e,x)\\ \text{Recipient}(e,y)\\ \text{Theme}(e,z)\\\hline\end{array};(mp@e)))))$$

Figure 2: Semantics of the verb *give*, ignoring tense. The core consists of a DRS introducing an event discourse referent e and relating it to the arguments using thematic role relations. The modifier semantics is applied to e, the result is combined with the core DRS using merge reduction (;). The rest is largely for dealing with NP arguments.

categories that determine the ways in which constituents are allowed to combine, right of it are semantic categories that specify the semantics of a constituent based on the semantics of the constituents that combined to form it. There are a number of combinatory rules besides forward application, and each specifies syntax and semantics in a similar way. Boxer's most important task is thus to assign a semantics to each word; given the derivation, the semantics of each phrase then follows deterministically.

The semantics of all functor categories, such as verbs and prepositions, must be prepared to combine with their arguments and therefore must be functions. They are specified using the notation of the lambda calculus. For example, the semantics of a ditransitive verb has the form $\lambda io.\lambda do.\lambda su.\mathbf{sent}$ where *io*, *do*, *su* are variables standing for the semantics of the indirect object, the direct object and the subject respectively, reflecting the order in which the verb arguments combine syntactically, and $\mathbf{sent}$ is the semantics of the resulting sentence, specified in terms of the verb meaning and io, do, su.

Boxer uses a neo-Davidsonian event semantics, expressing verb meanings using event individuals that permit an unlimited number of verb modifiers. This motivated a decision to represent sentence meanings themselves as functions—from event properties to sentence meanings—to allow modifiers to combine with them. $\mathbf{sent}$ above is therefore always of the form $\lambda mp.\mathbf{drs}$ where *mp* is a variable standing for an event property and $\mathbf{drs}$ is an expression that evaluates to a proper truth-theoretic meaning representation in the form of a discourse representation structure (DRS; Kamp and Reyle, 1993).

All noun phrases are analyzed as generalized quantifiers, i.e. functions from individual properties to DRSs. Figure 1 shows the semantics of *a member* and *every member* as examples. Verb semantics take the status of NPs as generalized quantifiers into account by using *predicate abstraction*: they apply an NP semantics to a property that expresses the core verb meaning, modifier meanings, and meanings of any previously absorbed arguments. For example, for a ditransitive verb, the $\mathbf{drs}$ part above has the form $(su@\lambda x.(do@\lambda y.(io@\lambda z.\mathbf{drs}')))$. Figure 2 shows a full semantics of the verb *give*.

3 Scope-mediating Categories

3.1 Verbs

It is easy to see from Figure 2 that in our lexical-semantic framework, it is the semantics of a verb that determines the relative scope of its arguments. For example, with the given lexical entry, the sentence *A lecturer gave every student a book* receives an interpretation with a unique lecturer, but if we swap the substrings $su@\lambda x$ and $io@\lambda y$, every student got a book from a potentially distinct lecturer. We thus account for scope ambiguity by giving verbs different interpretations, as in the proposal of Hendriks (1993). Boxer can generate an appropriate lexical entry for any scope order given a tag on the verb encoding the scope order of arguments—e.g., for a three-place verb, one of 312, 321, 132, 123, 231, 213.

Table 1: Eight types of scope interaction involving universal quantifiers, mediated by a preposition. Columns indicate what the preposition attaches to (verb or noun phrase), which of its arguments contains the universal quantifier (object or the phrase it attaches to), and whether the universal quantifier has wide or narrow scope. All examples are from the Groningen Meaning Bank (GMB); numbers in brackets are GMB document identifiers.

	att	$\forall$ in	$\forall$ scope	example	count
(a)	NP	obj	wide	Finally the gorgeous jewel of the order, gleaming upon *the breast* **of** *every member*, suggested "your Badgesty," which was adopted, and the order became popularly known as the Kings of Catarrh. [72/0696]	115
(b)	NP	att	wide	*All such attacks* **by** *drone aircraft* are believed to be carried out by U.S. forces. [76/0357]	32
(c)	VP	obj	wide	Jobs *grew* **in** *every sector except manufacturing*, with much of the growth due to hurricane clean-up efforts in Florida. [97/0059]	57
(d)	VP	att	wide	NATO says militants surrounded the outpost, *firing from all directions* **with** *rocket-propelled grenades, small arms and mortars*. [92/0311]	25
(e)	NP	obj	narrow	He is the former director of national intelligence, *the head* **of** *all U.S. intelligence agencies*. [59/0286]	44
(f)	NP	att	narrow	The official Xinhua news agency says *all 28 workers* **in** *a mine in northwestern Shaanxi province* died when an underground cable caught fire on Saturday night. [40/0608]	16
(g)	VP	obj	narrow	Responsibility for Africa is *currently fractured* **under** *all three*. [90/0450]	11
(h)	VP	att	narrow	Opening batsman Jamie How *led all scorers* **with** *88 runs* as New Zealand reached 203-4 in 42.1 overs. [13/0199]	17
(i)	NP	obj	neutral	He said methods such as abortion do not fight poverty or help a country's development but actually constitute "*the destruction* **of** *the poorest of all human beings*." [13/0428]	1
(j)	NP	att	neutral	It tacitly encouraged Iraq's minority Sunni Muslims to vote, saying *all segments* **of** *the Iraqi people* must go to the polls. [52/0038]	83
(k)	VP	obj	neutral	The preferred stock, which would have a dividend rate of $ 1.76 a year, would *be convertible into Heritage common* **at** *a rate of four common shares for each preferred*. [38/0686]	3
(l)	VP	att	neutral	The program airs in 40 countries worldwide, and *every Idolwinner records* **through** *Sony BMG*. [75/0494]	52

3.2 Prepositions

Scope interactions involving universal quantifiers and mediated by prepositions can roughly be classified along three binary distinctions: whether the preposition attaches to a verb or a noun, whether it is the object of the preposition or the phrase it attaches to that contains the universal quantifier, and whether the universal quantifier takes wide scope over the other argument of the preposition. Table 1 has an example for each situation.

Whether the universal quantifier must take wide scope depends on the desired interpretation. It should take wide scope if the non-universal argument introduces an entity that is non-specific, i.e. whose extension depends on the instantiation of the universally quantified-over variable. In this case, the universal argument should take wide scope. This is the case in (a–d). On the other hand, if the non-universal argument introduces an existentially quantified-over variable that is necessarily the same for all instantiations of the universally quantified-over variable, the non-universal argument must take wide scope. This is the case in (e–h). There are also cases were neither of the two criteria applies, e.g. because the non-universal argument is a definite that should be interpreted at a global level. Examples are shown

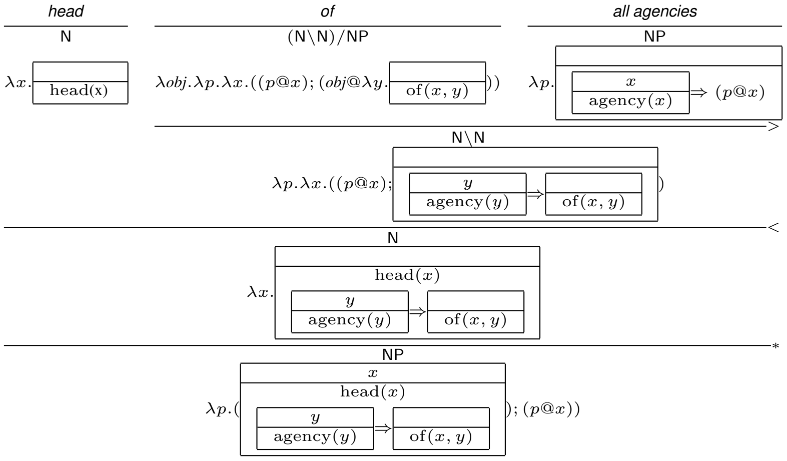


Figure 3: NP derivation with narrow-scope modifier

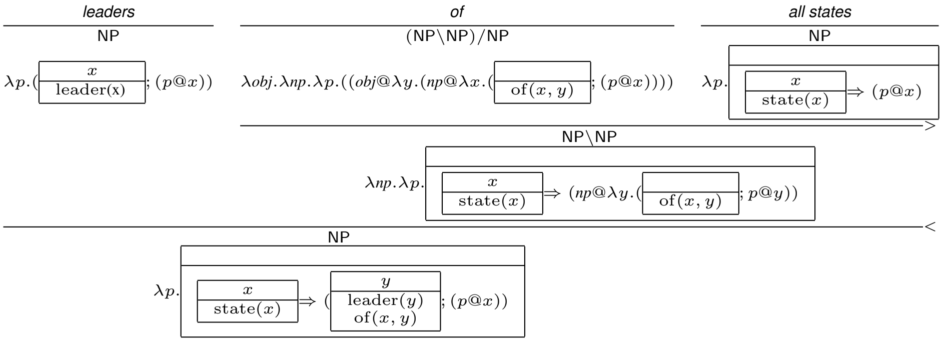


Figure 4: NP derivation with wide-scope modifier

in (i–l). Finally, in rare cases, both of the aforementioned criteria may apply simultaneously. The only example we found of this in the corpus is given in (1) below, where the modifier *in all major cities and towns* should be outscoped by *an Islamic alliance*, but outscope *members*. In such cases, our scheme is too coarse-grained to derive the desired semantics.

(1) *Members of an Islamic alliance and other parties took to the streets Saturday* **in** *all major cities and towns*, where speakers denounced General Musharraf for breaking his promise. [50/0210]

Scope alternation in these configurations can be effected using just two scope tags on prepositions, indicating modifier wide scope (inv as in a,c,f,h) vs. modified constituent wide scope (noninv as in b,d,e,g). For cases like (i–l), one tag can be used as the default.

In the case of verb phrase modifiers, preposition semantics effecting the desired scope behavior are straightforward to implement (Figure 5). For noun phrase modifiers, things are trickier. Firstly, narrow-scope modifiers need to apply to individual properties whereas wide-scope modifiers need to apply to generalized quantifiers. The easiest way to ensure this is to give the preposition the syntactic type $((N\backslash N)/NP)$ in the former case and $((NP\backslash NP)/NP)$ in the latter before parsing. Thus, in this case, the syntactic structure is affected. Both categories are well supported in the standard model of the C&C parser trained on CCGbank (Hockenmaier and Steedman, 2007), and although the assignment of lexical categories is integrated with parsing, the parser can be forced using tricks to pick a particular category. Example derivations for both cases are shown in Figures 3 and 4.

The second issue with noun phrase modifiers is more severe: specific modifiers of universally quantified noun phrases as in (f) cannot be accounted for in a similar way without heavily modifying the lexical-semantic framework because universally quantified noun phrases keep only their nuclear scope open for modification, not their restrictor. We will return to this limitation in the discussion.

$$\lambda obj.\lambda vp.\lambda subj.\lambda mp.((vp@subj)@\lambda e.(obj@\lambda y.(\begin{array}{|c|}\hline \\ \hline \text{with(e, y)} \\ \hline \end{array}; (mp@e))))$$

$$\lambda obj.\lambda vp.\lambda subj.\lambda mp.(obj@\lambda y.((vp@subj)@\lambda e.(\begin{array}{|c|}\hline \\ \hline \text{with(e, y)} \\ \hline \end{array}; (mp@e))))$$

Figure 5: Two possible semantics of a preposition attaching to a VP. Syntactically, it combines first with its object, then with the VP, then the resulting extended VP combines with the subject, hence $\lambda obj.\lambda vp.\lambda subj....$ Semantically, in the first entry, *vp* is first applied to *subj*, yielding a sentence meaning, which is then modified by applying it to an event property ($\lambda e....$). In the second entry, application to the preposition object to the corresponding property is moved out of the scope of all of this to give it scope over the verb and its arguments. In both cases, the additional λmp ensures that a proper sentence meaning results, allowing for additional modifiers.

4 Annotation and Results

For studying quantifier scope interactions in the wild, we use the Groningen Meaning Bank (GMB), a freely available annotated English corpus of currently a little over 1 million tokens in 7,600 documents, made up mainly of political news, country descriptions, fables and legal text. Since a pilot study showed that clear deviations from the default scope order of verbal arguments (subject $>$ objects in surface order) is very rare (only 12 of 206 cases), we decided to focus on scope interactions mediated by prepositions.

Using the syntactic annotation of the GMB we extracted all prepositions where either the object or the modified constituent contains one of the universally-quantifying determiners *every*, *each* and *all*. We discarded prepositions serving as discourse connectives as in ***With** all ballots cast, people are awaiting the results* because they relate propositions to each other rather than individuals with which we are concerned. We also discarded prepositions that are part of fixed expressions such as ***in** all*, ***at** all*, as well as the *of* in NPs such as *all **of** the leaders*, which we assume to mean the same as *all leaders*. Finally, we cast aside the prepositions *including*, *excluding* and *except* as deserving special treatment not within the scope of this work. This left us with 456 instances which were manually annotated by one annotator for scope behavior. The counts are shown in the last column of Table 1.

In cases of doubt, we generally preferred annotations resulting in the logically weaker reading. For example, in (2), we could either assume a separate termination event for each banking license or a single termination event for them all, and we preferred the former by giving the universal quantifier wide scope.

(2) the International Banking Repeal Act of 2002 resulted in *the termination* **of** *all offshore banking licenses* [03/0688]

5 Discussion

Our approach deals with scope alternation at the level of functor semantics, thus is syntactically constrained. There are some scope orderings, such as interleaved verb arguments and modifiers, that it cannot derive. Also, specific indefinites pose problems in cases such as (f), and will probably show need for special treatment even more clearly once we broaden our attention to their interactions with intensional verbs and negation, see e.g. Geurts (2010). A more powerful account of indefinites is proposed within CCG in Steedman (2012). However, it remains to be seen whether it can be implemented efficiently within a wide-coverage parser, see Kartsaklis (2010). Another phenomenon we haven't dealt with yet is scope interaction between universals and negation.

Nevertheless, our approach is empirically successful at accounting for 96% of all cases in some of the most common syntactic configurations giving rise to scope ambiguities involving universal quantifiers, namely modification of NPs and VPs by preposition phrases. It is also a straightforward extension to an existing semantic parsing system given an annotation of the input text with a layer of scope tags. In future work, we plan to provide this layer automatically by adapting techniques for the statistical resolution of scope, such as that of Manshadi and Allen (2011).

References

Andrew, G. and B. MacCartney (2004). Statistical resolution of scope ambiguity in natural language. Unpublished manuscript.

Basile, V., J. Bos, K. Evang, and N. Venhuizen (2012). Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, pp. 3196–3200.

Bos, J. (1996). Predicate Logic Unplugged. In P. Dekker and M. Stokhof (Eds.), *Proceedings of the Tenth Amsterdam Colloquium*, ILLC/Dept. of Philosophy, University of Amsterdam, pp. 133–143.

Bos, J. (2009). Towards a large-scale formal semantic lexicon for text processing. In C. Chiarcos, R. Eckart de Castilho, and M. Stede (Eds.), *Proceedings of the Biennal GSCL Conference 2009*, pp. 3–14. Gunter Narr Verlag.

Copestake, A., D. Flickinger, I. Sag, and C. Pollard (2005). Minimal recursion semantics: An introduction. *Journal of Research on Language and Computation 3*(2–3), 281–332.

Curran, J., S. Clark, and J. Bos (2007). Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, pp. 33–36.

Egg, M., A. Koller, and J. Niehren (2001). The constraint language for lambda structures. *Logic, Language and Information 10*, 457–485.

Geurts, B. (2010). Specific indefinites, presuppositions, and scope. In R. Buerle, U. Reyle, and T. E. Zimmermann (Eds.), *Presuppositions and discourse*, pp. 125–158. Emerald Group.

Hendriks, H. (1993). *Studied Flexibility: Categories and Types in Syntax and Semantics*. Ph. D. thesis, University of Amsterdam.

Higgins, D. and J. Sadock (2003). A machine learning approach to modeling scope preferences. *Computational Linguistics 29*(1), 73–96.

Hockenmaier, J. and M. Steedman (2007). Ccgbank: a corpus of ccg derivations and dependency structures extracted from the penn treebank. *Computational Linguistics 33*(3), 355–396.

Kamp, H. and U. Reyle (1993). *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Dordrecht: Kluwer.

Kartsaklis, D. (2010). Wide-coverage CCG parsing with quantifier scope. Master's thesis, University of Edinburgh.

Manshadi, M. and J. Allen (2011). Unrestricted quantifier scope disambiguation. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, Portland, Oregon, pp. 51–59. Association for Computational Linguistics.

Reyle, U. (1993). Dealing with ambiguities by underspecification: Construction, representation and deduction. *Journal of Semantics 10*, 123–179.

Srinivasan, P. and A. Yates (2009). Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pp. 1465–1474. Association for Computational Linguistics.

Steedman, M. (2001). *The Syntactic Process*. The MIT Press.

Steedman, M. (2012). *Taking Scope*. The MIT Press.

What is in a text, what isn't, and what this has to do with lexical semantics

Aurelie Herbelot
Universität Potsdam
aurelie.herbelot@cantab.net

Abstract

This paper queries which aspects of lexical semantics can reasonably be expected to be modelled by corpus-based theories such as distributional semantics or techniques such as ontology extraction. We argue that a full lexical semantics theory must take into account the *extensional* potential of words. We investigate to which extent corpora provide the necessary data to model this information and suggest that it may be partly learnable from text-based distributions, partly inferred from annotated data, using the insight that a concept's features are extensionally interdependent.

1 Introduction

Much work in computational linguistics relies on the use of corpora as evidential data, both for investigating language and for 'learning' about the world. Indeed, it is possible to inspect a great variety of phenomena, and get access to a lot of world knowledge, simply by having a large amount of text available. The purpose of this paper is to both acknowledge corpora as an invaluable data source for computational linguistics and point at their shortcomings. In particular, we want to argue that an appropriate representation of lexical meaning requires information beyond what is provided by written text and that this can be problematic for lexical models which rely entirely on corpora. Specifically, we wish to highlight how corpora fail to supply the information necessary to represent the *extension* of a term. In a nutshell, our argument is that the aspect of meaning represented by model theory is, in the best case, hard to extract and in the worst case not available at all when looking at corpus data. Building on this first discussion, we show that a concept's features are *extensionally* interdependent and, using this insight, propose that the part of model theory dealing with set relations (how much of set X is included in set Y?) may be learnt by exploiting a mixture of annotated (non-textual) data and standard distributional semantics. We present preliminary experiments investigating this hypothesis.

2 Text and extension

1. My cat – a mammal – likes sitting on the sofa.
2. There are more than two trees in this forest.
3. Cats have two eyes.

Sentences like 1–3 are not often found in corpora, or any kind of speech, for that matter. This is fortunate from the point of view of communication. In all three examples, the Gricean maxim of quantity is violated: 1 repeats information which is already implicitely given by the encyclopedic definition of the lexical item *cat*, while the other two express something obvious to anyone who has been in contact with forests/cats in their life (or indeed, who has read the encyclopedic definitions of *forest*/*cat*).

From the point of view of standard computational linguistics, this state of affairs is however undesirable. We will consider here two major approaches to extracting 'conceptual' knowledge from text:

distributional semantics and ontology extraction. Distributional semantics accounts for the lexical meaning of words by modelling, via appropriate weighting, their co-occurrence with other words (or any larger lexical context). The representation of a target word is thus a vector in a space where each dimension corresponds to a possible context. (Curran, 2003 and Turney and Pantel, 2010 are good overviews of this line of work). Ontology extraction, on the other hand, retrieves facts from text by searching for specific patterns: for instance, it is possible to find out that 'cat' is a hyponym of 'feline' by examining lexical patterns of the type 'X such as Y' (Hearst, 1992).

It should be clear that neither method can extract information which is simply absent from text. 3, for instance, may never occur in a corpus. This issue has been noted before, in particular in the psycholinguistics literature (Devereux et al., 2009; Andrews et al., 2009). To compensate, the use of manually obtained 'feature norms' is often advocated to complete corpus-based representations of concepts (e.g. a human might associate *bear* with the sensory-motor norms *claw, big, brown*, etc). But missing features are not the only issue. Let us return to 1. If we mentioned the word *mammal* everytime we speak of a cat, distributional semantics systems would be able to appropriately reflect the hyponymic relation between the two concepts – which in turn translates into an inclusion relation in model theory. But precisely because of the strong association between them, we (as good Griceans) are careful to leave the mammal out of the cat. Consequently, it is usual for a distribution to assign fairly low weights to features which we would argue are essential to the lexical representation of a word (see Geffet and Dagan, 2005 or Baroni et al., 2012 for attempts to capture the inclusion relation despite the flaws in the data).

Pattern-based methods obviously fare better when attempting to extract standard lexical relations. They do not, however, provide a true *extensional* analysis of the information they retrieve. For instance, from the snippet *four-legged animals such as cats and dogs*, we should not conclude that the set of dogs is included in the set of four-legged animals – a fair amount of dogs only have three legs. Our point is that relations obtained from corpora (or by coding feature norms) are essentially intensional: they do not model the part of semantics dealing with set relations and thus do not reflect our *expectations* with regard to a particular concept (i.e. how *likely* is it that a given bear is big, brown, etc?) We argue that such expectations are part of lexical semantics, as they mediate our use of words.[1] For instance, we strongly expect a dog to have four legs, but are not overly surprised when seeing a three-legged dog (the set of dogs is mostly, but not entirely, included in the set of four-legged things) and so would still call it a dog. Conversely, a forest with one tree is not intuitively a forest (the sets of trees making up each instance in the set of forests *all* have a 'large' cardinality).

In what follows we investigate how such set relations can be inferred from prior knowledge: we recall that features (whether of the distributional type or the sensory-motor type) are dependent on each other and that this dependency can be exploited to learn human-like expectations with regard to extension.

3 The extensional dependency hypothesis

Let us assume a particular kind of conceptual representation which consists of a vector of weighted features, as in distributional semantics, but where the weights are the *expected probabilities* of an instance of the concept to be associated with a particular distributional feature in the real world. So for instance, the feature *mammal* has a weight of 1 for the concept *cat* because all cats are mammals. In the same concept, *black* has perhaps a weight of 0.1 (assuming that one cat in ten is black). We call such representations **extensional distributions**, because each probability reflects some expectation about the inclusion relation between two sets.

Let us further assume a distributional space with n dimensions and let us refer to the extensional distribution of A as A$^\circ$. We hypothesise that the value of A$^\circ$ along a dimension d_k is dependent on the value of A$^\circ$ along all other dimensions $d_{1...n}$ in that space. Intuitively, this means that the probability that a cat (habitually) eats is dependent on the probability of that cat to (habitually) sleep, run, communicate, to be made of stone or to write books. In other words, the extensional distribution of a typical cat x

[1] Of course, expectations reflect certain common beliefs and no extensionally true facts about the world. But those beliefs can be modelled using the set-theory machinery, i.e. they can be regarded as a possible world.

reflects its status as a living (non-human) being, which in turn implies a high probability of cat$^\circ$ along the dimension *eat*. We call this the **extensional dependency hypothesis** (see Devereux et al., 2009 for related comments on this effect).

Such dependencies mean that learning the probabilities of certain features in relation to instances of a given target word can be greatly facilitated by learning another, information-rich feature. For instance, knowing that a is a bird allows us to infer many properties about a, e.g. that it lays eggs if it is female, that it probably flies, or that it perhaps builds nests.

In the absence of any supervision, it is hard to find out which features are most inference-producing. However, we can adopt a different strategy to learn inference rules. Let us assume that there *is* a correspondence between distributional data and the real world for at least some features. For example, we might find that as long as the predicate *be_v+fish_n* has been seen next to *pike* – be it only once –, we can infer that all pikes are fish. It is a property of the feature that if it applies to one individual of a kind, it will apply to all of them (contrast this with, e.g. *black_a*). If we can be reasonably certain that, given enough data, *be_v+fish_n* will be seen next to *pike*, we have a way to learn a real world probability from corpus data which we are confident in and which may be used for producing inferences. The rest of this paper investigates this hypothesis.

4 Experiments

4.1 A distributional system

Talking of '(expected) probabilities in the real world' has consequences in terms of choosing a particular notion of context for building our corpus-based distributions. Consider a system where context is defined by a word window around the target. *mouse* may be one of the terms sometimes appearing in the context of *cat*, but having a feature *mouse* does not tell us anything about how mice and cats are related in the real world (do mice eat cats? are made of cats? sell cats?) and so, the 'association' *mouse-cat* cannot be assigned a probability. In contrast, if we choose as context semantic dependencies of the type *_+eat+mouse* or *_+be+kept+as+pet*, where _ indicate the position of the target word, there is a clear interpretation of the context as being related to the target in the real world (what is the probability of a cat to eat mice? to be kept as a pet?) Consequently, we build a distributional system using dependency relations. Our data is the Wikiwoods corpus (Flickinger et al., 2010), a Wikipedia snapshot parsed with the English Resource Grammar (ERG), which we convert into a Dependency Minimal Recursion Semantics (DMRS, Copestake, 2009) format for our experiments. We only consider a number of dependency structures as context for a target word: adjectives, adverbs, prepositions and verbs, with their direct arguments, possessive constructs and coordination. The weight of a target word along a particular dimension of the semantic space is given by the normalised PMI measure proposed by Bouma (2007).

To speed up processing, we only retain dimensions which do occur with the lexical items in our experimental set (see 4.3 for a description). This results in our semantic space having 15799 dimensions.

4.2 The learning system

We design a system based on bootstrapping, which learns extensional distributions. The idea is to learn from our corpus the features which we are most confident in and use those to further the learning process.

Let us assume we wish to learn real-world probabilities for some features $F_1...F_n$, as applied to instances of various target words $w_k...w_m$ (F_1 might be *lay_v+egg_n* while w_k might be *aardvark*). Let us also assume that we have some annotated data which provides us with 'rough' probabilities for a number of feature-instance pairs. For convenience of annotation – and because we do not expect humans to have an accurate probabilistic model of the world –, we express those probabilities using the natural language quantifiers *no, few, some, most, all* (see Herbelot and Copestake, 2011 for related work on resolving underspecified quantification). So we might already know that *most* swallows migrate.

The first iteration of our bootstrapping process runs over each feature F in $\{F_1...F_n\}$. A machine learning algorithm is fed the corpus-based distributions of the training instances (let us call them

$w_1...w_j$), together with the probabilities $p_m(F|w_1...j)$ from the annotated data. A classifier is learnt for F and its precision estimated by performing leave-one-out cross-validation on the training data. The classifier with the best precision, say $C(F_i)$ is recorded, together with its decisions, $p(F_i|w_1...j)$. The feature F_i is considered 'learnt'.

From the second iterations on, the following process takes place. For each feature F which has not yet been learnt, the machine learning algorithm is fed the corpus-based distributions of $w_1...w_j$, together with the values of the learnt features $p(F_{learnt}|w_1...j)$ and the manually annotated probabilities $p_m(F|w_1...j)$. As before, a classifier is learnt and its precision calculated by performing leave-one-out cross-validation on the training data and the classifier with the best precision, as well as its decisions on the training data, are recorded for use in the next iteration.

When classifying new, unseen instances, the classifiers are applied to the data in the order they were learnt during the training process. As an example, let us assume that we have learnt the probabilities of the features *be_v+fish_n* and *aquatic_a* for the animals *aardvark, cod* and *dolphin*. Let us also assume that *be_v+fish_n* was learnt first (it was classified with higher precision than *aquatic_a* in the first iteration of the algorithm). Let us further suppose that in the second iteration, the initial classifier for *aquatic_a* was modified to take into account the learnt feature *be_v+fish_n* (i.e. the precision of the classifier was increased by using the new information). Presented with a new classification problem, say *salmon*, the system would first try to find out the probability *p(be_v+fish_n|salmon)* and then use that figure to estimate *p(aquatic_a|salmon)*.

4.3 The dataset

We attempt to learn the quantification of a number of animal-feature pairs. The animals and features used in our experiment are chosen as follows.

The animals are picked by selecting the entries in the Wikipedia 'List of animals'[2] which have an occurrence count over 2000 in our corpus. This results in a set of 72 animals.

The features are chosen manually amongst the 50 most highly weighted vector components of ten different animal distributions. The animals considered are selected semi-randomly: we make sure that the most common types are included (mammals, fish, insects, birds, invertebrates). Features which satisfy the following conditions are included in the experiment: a) the feature must be applicable to the animal at the 'individual' level, i.e. it cannot be a temporary state of the animal, or apply to the species collectively (*black_a* is appropriate while $wounded\_a$ or $endangered\_a$ are not) b) the feature must be semantically 'complete', i.e. make sense when applied to the animal in isolation (*be_v+mammal_n* is appropriate while *behaviour_n+of_p* is not).

The feature selection exercise results in 54 vector components being selected.

Given our 72 animals and 54 features, we ask a human annotator to mark each animal-feature pair with a 'probability', expressed as a quantifier. Possible values are *no, few, some, most, all.* The guidelines for the annotation task can be seen at `http://www.cl.cam.ac.uk/~ah433/material/iwcs13-annot.pdf`.

5 Results

The following is based on an implementation of our system using Weka's[3] C4.5 (also referred to as J48) classifier (Quinlan, 1993). The C4.5 classifier produces a decision tree which can be inspected and is therefore particularly appropriate to 'check' learnt rules against human intuition. We perform leave-one-out validation on the first 10 animals in our dataset (for instance, we predict the values of our 54 features for *ant* using a training set consisting of all other animals). For each animal, the system therefore runs 1485 iterations of the classifier.

[2] `http://en.wikipedia.org/wiki/List_of_animals#Animals_.28by_common_name.29`

[3] `http://http://www.cs.waikato.ac.nz/~ml/weka/`

5.1 Finding relationships between corpora and the real world

In order to investigate whether real-world probabilities can be derived from corpus data, we first run a baseline system which classifies animal-feature pairs into our five quantifier classes, *no, few, some, most, all*, using only the corpus-based distributions of the animals in our training set.

The overall precision of the baseline system over our 540 predicted values is 0.51. As hypothesised, some features are learnt with high precision using word distributions only: *be_v+bird_n* is one such example, classifying all but one instances correctly (its estimated performance on the training data, as per cross-validation, is 0.95). Others, such as *carnivorous_a* consistently receive an incorrect classification (precision of 0 on the test data and 0.24 on the training data).

5.2 Learning inferences

We now turn to the question of learning rules which reflect real-world relationships. Because of space constraints, we are unable to reproduce actual output from the system, but examples can be seen at `http://www.cl.cam.ac.uk/~ah433/material/iwcs13-examples.pdf`. As illustration, we consider here the classifiers produced for the feature *aquatic_a*.

The baseline classifier (the one learnt from distributional data only) does use distributional features that are associated with water (*mediterranean_a*, *in_p()+water_n*), but overall, the decision tree is rather far from the way we might expect a human to attribute the feature *aquatic_a* to an animal species, and mostly includes seemingly irrelevant features such as *variant_of* or again *fascinating*.

The final classifier produced by our learning system, in contrast, involves the learnt probability for the feature *terrestrial_a*. It captures the facts that a) non-terrestrial species are aquatic and b) terrestrial species which live near water can also be said to be aquatic (cf. otters or beavers).

Other examples of real world dependencies learnt in the iterative process include *lay_v+egg_n* and *woolly_a* being dependent on *be_v+mammal_n*, *have_v+hair_n* depending on *woolly_a* and *be_v+insect_n*, or again *walk_v* depending on *be_v+mammal_n* and *fly_v*.

5.3 Overall performance

The overall precision of the iterative system is 0.48, which makes it lag behind the baseline by 3%. The result is disappointing but can be explained when inspecting the output. The strength of the system in catching meaningful dependencies between features, which we illustrated in the previous section, is also what makes it lose accuracy on our small dataset. Indeed, the dependencies integrated in the final classifiers assume that the learnt features were correctly predicted for the animal under consideration. This is unfortunately not the case, and errors accumulate during bootstrapping. For instance, when running on the test instance *ape*, the system classifies the features *woolly_a* and *lay_v+egg_n* by relying on the prediction for *be_v+mammal_n*. The precision of the system on that feature, however, is only 0.5. In order to truly evaluate the performance of the system, we suggest experimenting with a larger dataset.

6 Conclusion

This paper argued that there is a part of lexical semantics which is not dealt with by modelling techniques relying entirely on corpora. Indeed, standard speech tends to omit information which is true from an extensional point of view but irrelevant for successful communication. In order to retrieve the extensional potential of lexical terms, we proposed a separate, distribution-like representation which provides the *expected* real-world probabilities of instance-feature pairs and which we refer to as extensional distribution. We argued that a) for *some* features, there is a correspondence between corpus data and real world which can be learnt b) probabilities in the extensional distribution are dependent on each other and it is possible to infer unknown values from already learnt ones (including those learnt from corpus data).

Acknowledgement

The author is in receipt of a postdoctoral fellowship from the Alexander von Humboldt foundation.

References

Andrews, M., G. Vigliocco, and D. Vinson (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review 116*(3), 463.

Baroni, M., R. Bernardi, N.-Q. Do, and C.-c. Shan (2012). Entailment above the word level in distributional semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL12).*

Bouma, G. (2007). Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of GSCL*, 31–40.

Copestake, A. (2009). Slacker semantics : why superficiality , dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL09)*, Athens, Greece, pp. 1–9.

Curran, J. (2003). *From Distributional to Semantic Similarity*. Ph. D. thesis.

Devereux, B., N. Pilkington, T. Poibeau, and A. Korhonen (2009). Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data. *Research on Language and Computation 7*, 137.

Flickinger, D., S. Oepen, and G. Ytrestol (2010). Wikiwoods: Syntacto-semantic annotation for english wikipedia. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC10).*

Geffet, M. and I. Dagan (2005). The distributional inclusion hypothesises and lexical entailment. In *Proceedings Of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 107–114.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. Volume 14th International Conference on Computational Linguistics (COLING 92), pp. 539.

Herbelot, A. and A. Copestake (2011). Formalising and specifying underquantification. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, Oxford, United Kingdom, pp. 165–174.

Quinlan, J. R. (1993). *Programs for Machine Learning*. Morgan Kaufman.

Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research 37*, 141–188.

Semantic Similarity Computation for Abstract and Concrete Nouns Using Network-based Distributional Semantic Models

Elias Iosif*, Alexandros Potamianos*, Maria Giannoudaki*, Kalliopi Zervanou†
* Dept. of Electronics & Computer Engineering, Technical University of Crete, Greece
{iosife,potam,maria}@telecom.tuc.gr
†Centre for Language Studies, Radboud University Nijmegen, The Netherlands
k.zervanou@let.ru.nl

Abstract

Motivated by cognitive lexical models, network-based distributional semantic models (DSMs) were proposed in [Iosif and Potamianos (2013)] and were shown to achieve state-of-the-art performance on semantic similarity tasks. Based on evidence for cognitive organization of concepts based on degree of concreteness, we investigate the performance and organization of network DSMs for abstract vs. concrete nouns. Results show a "concreteness effect" for semantic similarity estimation. Network DSMs that implement the maximum sense similarity assumption perform best for concrete nouns, while attributional network DSMs perform best for abstract nouns. The performance of metrics is evaluated against human similarity ratings on an English and a Greek corpus.

1 Introduction

Semantic similarity is the building block for numerous applications of natural language processing (NLP), such as grammar induction [Meng and Siu (2002)] and affective text categorization [Malandrakis et al. (2011)]. Distributional semantic models (DSMs) [Baroni and Lenci (2010)] are based on the distributional hypothesis of meaning [Harris (1954)] assuming that semantic similarity between words is a function of the overlap of their linguistic contexts. DSMs are typically constructed from co-occurrence statistics of word tuples that are extracted from a text corpus or from data harvested from the web. A wide range of contextual features are also used by DSM exploiting lexical, syntactic, semantic, and pragmatic information. DSMs have been successfully applied to the problem of semantic similarity computation. Recently [Iosif and Potamianos (2013)] proposed *network-based DSMs* motivated by the organization of words, attributes and concepts in human cognition. The proposed semantic networks can operate under either the *attributional similarity* or the *maximum sense similarity* assumptions of lexical semantics. According to attributional similarity [Turney (2006)], semantic similarity between words is based on the commonality of their sense attributes. Following the maximum sense similarity hypothesis, the semantic similarity of two words can be estimated as the similarity of their two closest senses [Resnik (1995)]. Network-based DSMs have been shown to achieve state-of-the-art performance for semantic similarity tasks.

Typically, the *degree of semantic concreteness* of a word is not taken into account in distributional models. However, evidence from neuro- and psycho-linguistics demonstrates significant differences in the cognitive organization of abstract and concrete nouns. For example, Kiehl et al. (1999) and Noppeney and Price (2004) show that concrete concepts are processed more efficiently than abstract ones (aka "the concreteness effect"), i.e., participants in lexical decision tasks recall concrete stimuli faster than abstract. According to dual code theory [Paivio (1971)], the stored semantic information for concrete concepts is both verbal and visual, while for abstract concepts stored information is only verbal. Neuropsychological studies show that people with acquired dyslexia (deep dyslexia) face problems in reading abstract nouns aloud [Coltheart (2000)], *verifying that concrete and abstract concepts are stored in different regions of the human brain anatomy* [Kiehl et al. (1999)]. The reversal concreteness effect is also reported for people with semantic dementia with a striking impairment in semantic memory [Papagno et al. (2009)].

Motivated by this evidence, we study the semantic network organization and performance of DSMs for estimating the semantic similarity of abstract vs. concrete nouns. Specifically, we investigate the validity of the maximum sense and attributional similarity assumptions in network-based DSMs for abstract and concrete nouns (for both English and Greek).

2 Related Work

Semantic similarity metrics can be divided into two broad categories: (i) metrics that rely on knowledge resources, and (ii) corpus-based metrics. A representative example of the first category are metrics that exploit the WordNet ontology [Miller (1990)]. Corpus-based metrics are formalized as DSM [Baroni and Lenci (2010)] and are based on the distributional hypothesis of meaning [Harris (1954)]. DSM can be categorized into unstructured (unsupervised) that employ a bag-of-words model [Agirre et al. (2009)] and structured that rely on syntactic relationships between words [Pado and Lapata (2007)]. Recently, motivated by the graph theory, several aspects of the human languages have been modeled using network-based methods. In [Mihalcea and Radev (2011)], an overview of network-based approaches is presented for a number of NLP problems. Different types of language units can be regarded as vertices of such networks, spanning from single words to sentences. Typically, network edges represent the relations of such units capturing phenomena such as co-occurrence, syntactic dependencies, and lexical similarity. An example of a large co-occurrence network is presented in [Widdows and Dorow (2002)] for the automatic creation of semantic classes. In [Iosif and Potamianos (2013)], a new paradigm for implementing DSMs is proposed: a two tier system in which corpus statistics are parsimoniously encoded in a network, while the task of similarity computation is shifted (from corpus-based techniques) to operations over network neighborhoods.

3 Corpus-Based Baseline Similarity Metrics

Co-occurrence-based: The underlying assumption of co-occurrence-based metrics is that the co-existence of words in a specified contextual environment indicates semantic relatedness. In this work, we employ a widely-used co-occurrence-based metric, namely, Dice coefficient [Iosif and Potamianos (2010)]. The Dice coefficient between words w_i and w_j is defined as follows: $D(w_i, w_j) = \frac{2f(w_i,w_j)}{f(w_i)+f(w_j)}$, where $f(.)$ denotes the frequency of word occurrence/co-occurrence. Here, the word co-occurrence is considered at the sentential level, while D can be also defined with respect to broader contextual environments, e.g., at the paragraph level [Véronis (2004)].
Context-based: The fundamental assumption behind context-based metrics is that *similarity of context implies similarity of meaning* [Harris (1954)]. A contextual window of size $2H + 1$ words is centered on the word of interest w_i and lexical features are extracted. For every instance of w_i in the corpus the H words left and right of w_i formulate a feature vector v_i. For a given value of H the context-based semantic similarity between two words, w_i and w_j, is computed as the cosine of their feature vectors: $Q^H(w_i, w_j) = \frac{v_i.v_j}{||v_i||\ ||v_j||}$. The elements of feature vectors can be weighted according various schemes [Iosif and Potamianos (2010)], while, here we use a binary scheme.

4 Network-based Distributional Semantic Models

Here, we summarize the main ideas of network-based DSMs as proposed in [Iosif and Potamianos (2013)]. The network is defined as an undirected (under a symmetric similarity metric) graph $F = (V, E)$ whose the set of vertices V are all words in our lexicon L, and the set of edges E contains the links between the vertices. The links (edges) between words in the network are determined and weighted according to the pairwise semantic similarity of the vertices. The network is a parsimonious representation of corpus statistics as they pertain to the estimation of semantic similarities between word-pairs in the lexicon. In addition, the network can be used to *discover relations that are not directly observable in the data*; such relations emerge via the systematic covariation of similarity metrics. For each word (reference word) that is included in the lexicon, $w_i \in L$, we consider a sub-graph of F, $F_i = (N_i, E_i)$, where the set of vertices N_i includes in total n members of L, which are linked with w_i via edges E_i. The F_i sub-graph is referred to as the semantic neighborhood of w_i. The members of N_i (neighbors of w_i) are selected according to a semantic similarity metric (co-occurrence-based D or context-based Q^H defined in Section 3) with respect to w_i, i.e., the n most similar words to w_i are selected. Next, we present two semantic similarity metrics that utilize the notion of semantic neighborhood [Iosif and Potamianos (2013)].

4.1 Maximum Similarity of Neighborhoods

This metric is based on the hypothesis that the similarity of two words, w_i and w_j, can be estimated by *the maximum similarity of their respective sets of neighbors*, defined as follows:

$$M_n(w_i, w_j) = \max\{\alpha_{ij}, \alpha_{ji}\}, \quad \text{where } \alpha_{ij} = \max_{x \,\in\, N_j} S(w_i, x), \;\; \alpha_{ji} = \max_{y \,\in\, N_i} S(w_j, y). \tag{1}$$

α_{ij} (or α_{ji}) denotes the maximum similarity between w_i (or w_j) and the neighbors of w_j (or w_i) that is computed according to a similarity metric S: in this work either D or Q^H. N_i and N_j are the set of neighbors for w_i and w_j, respectively. The definition of M_n is motivated by the maximum sense similarity assumption. Here the underlying assumption is that the most salient information in the neighbors of a word are semantic features denoting senses of this word.

4.2 Attributional Neighborhood Similarity

The similarity between w_i and w_j is defined as follows:

$$R_n(w_i, w_j) = \max\{\beta_{ij}, \beta_{ji}\}, \quad \text{where } \beta_{ij} = \rho(C_i^{N_i}, C_j^{N_i}), \;\; \beta_{ji} = \rho(C_i^{N_j}, C_j^{N_j}) \tag{2}$$

$$\text{where } C_i^{N_i} = (S(w_i, x_1), S(w_i, x_2), \ldots, S(w_i, x_n)), \quad \text{and } N_i = \{x_1, x_2, \ldots, x_n\}.$$

Note that $C_j^{N_i}$, $C_i^{N_j}$, and $C_j^{N_j}$ are defined similarly as $C_i^{N_i}$. The ρ function stands for the Pearson's correlation coefficient, N_i is the set of neighbors of word w_i, and S is a similarity metric (D or Q^H). Here, we aim to exploit the entire semantic neighborhoods for the computation of semantic similarity, as opposed to M_n where a single neighbor is utilized. The motivation behind this metric is attributional similarity, i.e., we assume that semantic neighborhoods encode attributes (or features) of a word. Neighborhood correlation similarity in essence compares the distribution of semantic similarities of the two words on their semantic neighborhoods. The ρ function incorporates the covariation of the similarities of w_i and w_j with respect to the members of their semantic neighborhoods.

5 Experimental Procedure

Lexica and corpora creation: For English we used a lexicon consisting of $8,752$ English nouns taken from the SemCor3[1] corpus. In addition, this lexicon was translated into Greek using Google Translate[2], while it was further augmented resulting into a set of $9,324$ entries. For each noun an individual query was formulated and the $1,000$ top ranked results (document snippets) were retrieved using the Yahoo! search engine[3]. A corpus was created for each language by aggregating the snippets for all nouns of the lexicon.
Network creation: For each language the semantic neighborhoods of lexicon noun pairs were computed following the procedure described in Section 4 using either co-occurrence D or context-based $Q^{H=1}$ metrics [4].
Network-based similarity computation: For each language, the semantic similarity between noun pairs was computed applying either the max-sense M_n or the attributional R_n network-based metric. The underlying semantic similarity metric (the S metric in (1) and (2)) can be either D or Q^H. Given that for both neighborhood creation and network-based semantic similarity estimation we have the option of D or Q^H, a total of four combinations emerge for this two-phase process: (i) D/D, i.e., use co-occurence metric D for both neighborhood selection and network-based similarity estimation, (ii) D/Q^H, (iii) Q^H/D, and (iv) Q^H/Q^H.

6 Evaluation Datasets

The performance of network-based similarity metrics was evaluated for the task of semantic similarity between nouns. The Pearson's correlation coefficient was used as evaluation metric to compare estimated similarities against the ground truth (human ratings). The following datasets were used:
English (WS353): Subset of WS353 dataset [Finkelstein et al. (2002)] consisting of 272 noun pairs (that are also included in the SemCor3 corpus).
Greek (GIP): In total, 82 native speakers of modern Greek were asked to score the similarity of the noun pairs in a range from 0 (dissimilar) to 4 (similar). The resulting dataset consists of 99 nouns pairs (a subset of pairs translated from WS353) and is freely available [5].
Abstract vs. Concrete: From each of the above datasets two subsets of pairs were selected, where both nouns in the pair are either abstract or concrete, i.e., pairs consisting of one abstract and one concrete nouns were ruled out. More specifically, 74 abstract and 74 concrete noun pairs were selected from WS353, for a total of 148 pairs. Regarding GIP, 18 abstract and 18 concrete noun pairs were selected, for a total of 36 pairs.

[1] http://www.cse.unt.edu/~rada/downloads.html

[2] http://translate.google.com/

[3] http://www.yahoo.com//

[4] We have also experimented with other values of context window H not reported here for the sake of space. However, the highest performance was achieved for $H = 1$.

[5] http://www.telecom.tuc.gr/~iosife/downloads.html

7 Results

The performance of the two proposed network-based metrics, M_n and R_n, for neighborhood size of 100, is presented in Table 1 with respect to the English (WS353) and Greek (GIP) datasets. Baseline performance (i.e., no use of the network) is also shown for co-occurrence-based metric D and context-based metric Q^H. For the max-sense similarity $M_{n=100}$ metric, the use of the co-occurrence metric D for neighbor selection yields the best correlation performance for both languages. For the attributional similarity $R_{n=100}$ metric, best performance is achieved when using the context-based metric D for the selection of neighbors in the network. As explained in [Iosif and Potamianos (2013)], the neighborhoods selected by the D metrics tend to include words that denote word senses (yielding best results for similarity), while neighborhoods computed using the Q^H metric are semantically broader including word attributes (yielding best results for attributional similarity). The network-based DSM results are also significantly higher compared to the baseline for both languages. The best results achieved by D/Q^H for the $M_{n=100}$, and Q^H/D for the $R_{n=100}$ are consistent with the results reported in [Iosif and Potamianos (2013)] for English. The best performing metric for English is $M_{n=100}$ (max-sense) while for Greek $R_{n=100}$ (attributional). Overall, utilizing network neighborhoods for estimating semantic similarity can achieve good performance[6], and the type of metric (feature) used to select the neighborhood is a key performance factor.

Language: dataset	Number of pairs	Baseline		Network metric	Neighbor selection / Similarity computation			
		D	Q^H		D/D	D/Q^H	Q^H/D	Q^H/Q^H
English: WS353	272	0.22	0.30	$M_{n=100}$	**0.64**	**0.64**	0.47	0.46
				$R_{n=100}$	0.50	0.14	**0.56**	**0.57**
Greek: GIP	99	0.25	0.13	$M_{n=100}$	**0.51**	**0.51**	0.04	0.04
				$R_{n=100}$	-0.11	0.03	**0.66**	0.11

Table 1: Pearson correlation with human ratings for neighborhood-based metrics for English and Greek datasets. Four combinations of the co-occurrence-based metric D and the context-based metric Q^H were used for the definition of semantic neighborhoods and the computation of similarity scores. Baseline performance is also shown.

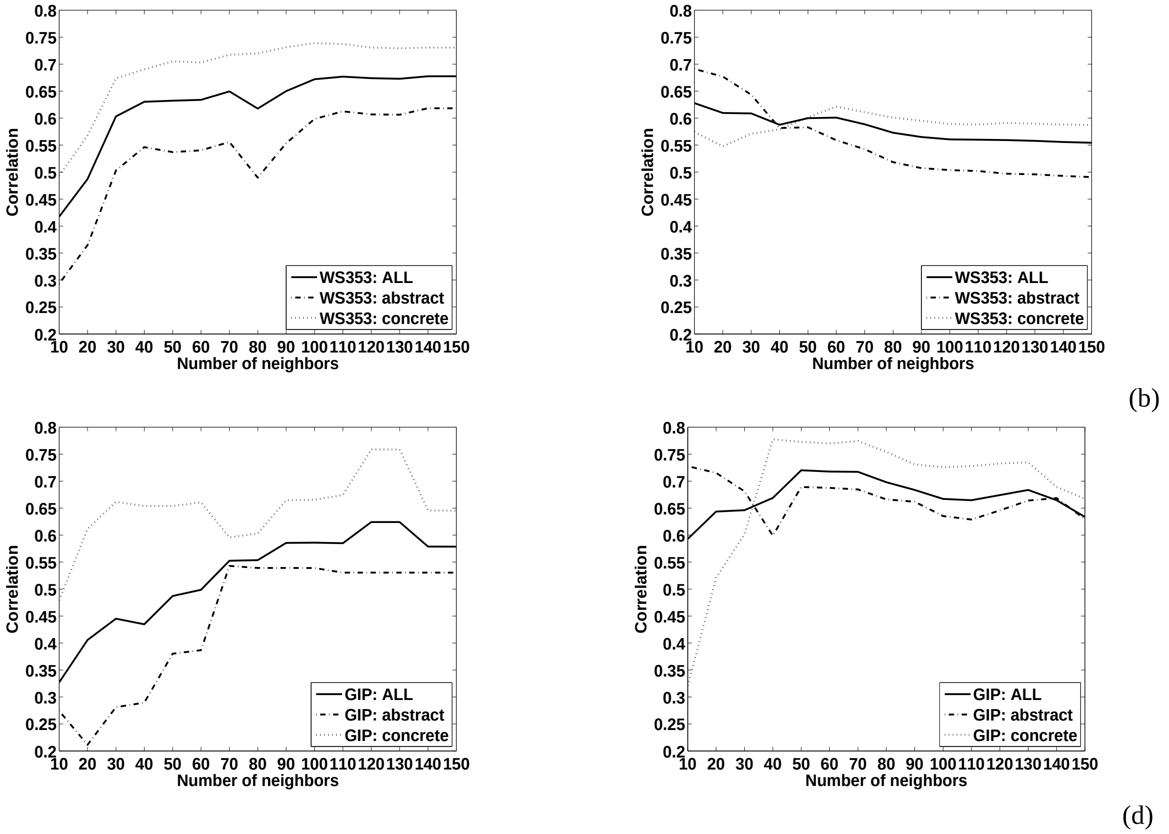
Next, we investigate the performance of the network metrics with respect to the neighborhood size n for the abstract and concrete noun pairs included in English and Greek datasets. The performance of the max-sense M_n (D/Q^H) metric is shown in Fig. 1(a),(c) for the (subsets of) WS353 and GIP, respectively. The performance over the whole (abstract and concrete) dataset is shown with a solid line. Similarly the results for the attributional R_n (Q^H/D) metric are shown in Fig. 1(b),(d). The main conclusions for these experiments (for both languages) are: 1) The correlation performance for concrete noun pairs is higher than for abstract noun pairs. 2) For concrete nouns the max-sense M_n metric achieves best performance, while for abstract nouns the attributional R_n metric is the top performer. 3) For the R_n network metric, very good performance is achieved for abstract noun pairs for a small neighborhood size n (around 10), while for concrete nouns larger neighborhoods are needed (up to 40 and 30 neighbors, for English and Greek, respectively).

Neighbor selection metric	Number of reference nouns	Type of reference nouns	Type of neighbors (abstract/concrete)			
			English (WS353)		Greek (GIP)	
			abstract	concrete	abstract	concrete
D	15	abstract	**76%**	24%	**82%**	18%
D	15	concrete	36%	**64%**	23%	**77%**
Q^H	15	abstract	**82%**	18%	**91%**	9%
Q^H	15	concrete	31%	**69%**	31%	**69%**

Table 2: Distribution of abstract vs. concrete nouns in (abstract/concrete noun) neighbourhoods.

In order to further investigate the network organization for abstract vs. concrete nouns, we manually inspected the top twenty neighbors of 30 randomly selected nouns (15 abstract and 15 concrete) and classified each neighbor as either abstract or concrete. The distributions of abstract/concrete neighbors are shown in Table 2 as a function of neighbor selection metric (D vs. Q^H) and reference noun category. It is clear, that the neighborhoods of abstract nouns contain mostly abstract concepts, especially for the Q^H neighbor selection metric (similarly the neighborhoods of concrete nouns contain mainly concrete concepts). The neighbors of concrete nouns mainly belong to the same semantic class (e.g., “vehicle”, “bus” for “car”) often corresponding to relevant senses. The

[6] The best correlation score for the WS353 dataset does not exceed the top performance (0.68) of unsupervised DSMs [Agirre et al. (2006)]. However, we have found that the proposed network metrics obtain state-of-the-art results for other standard datasets, e.g., 0.87 for [Rubenstein and Goodenough (1965)] and 0.91 for [Miller and Charles (1998)].



(a)

(b)

(c)

(d)

Figure 1: Correlation as a function of number of neighbors for network-based metrics. Max-sense M_n (D/Q^H) for datasets: (a) English and (c) Greek. Attributional R_n (Q^H/D) for datasets: (b) English and (d) Greek.

neighbors of the abstract nouns have an attributive function, reflecting relative attributes and/or aspects of the referent nouns (e.g., “religion”, “justice” for “morality”).

8 Discussion

We investigated the performance of network-based DSMs for semantic similarity estimation for abstract and concrete noun pairs of English and Greek. We observed a “concreteness effect”, i.e., performance for concrete nouns was better than for abstract noun pairs. The assumption of maximum sense similarity as encoded by the M_n metric consistently yielded higher performance for the case of concrete nouns, while the semantic *similarity of abstract nouns was better estimated via the attributional similarity assumption* as implemented by the R_n metric. The results are consistent with the initial hypothesis that differences in cognitive organization may warrant different network organization in DSMs. In addition, abstract concepts were best modeled using an attributional network DSM with small semantic neighborhoods. This is a first step towards the better understanding of the network organization of DSMs for different categories of concepts. In terms of computation algorithms of semantic similarity, it might prove advantageous to define a metric that combines the maximum sense and attributional assumptions based on the semantic concreteness of the words under investigation. Further research on more data and languages is needed to verify the universality of the findings.

9 Acknowledgements

Elias Iosif, Alexandros Potamianos, and Maria Giannoudaki were partially funded by the PortDial project (“Language Resources for Portable Multilingual Spoken Dialogue Systems”) supported by the EU Seventh Framework Programme (FP7), grant number 296170.

References

Agirre, E., E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 19–27.

Agirre, E., D. Martínez, O. L. de Lacalle, and A. Soroa (2006). Two graph-based algorithms for state-of-the-art WSD. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pp. 585–593.

Baroni, M. and A. Lenci (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics 36*(4), 673–721.

Coltheart, M. (2000). Deep dyslexia and right-hemisphere reading. *Brain and Language 71*, 299–309.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems 20*(1), 116–131.

Harris, Z. (1954). Distributional structure. *Word 10*(23), 146–162.

Iosif, E. and A. Potamianos (2010). Unsupervised semantic similarity computation between terms using web documents. *IEEE Transactions on Knowledge and Data Engineering 22*(11), 1637–1647.

Iosif, E. and A. Potamianos (2013). Similarity Computation Using Semantic Networks Created From Web-Harvested Data. *Natural Language Engineering (submitted)*.

Kiehl, K. A., P. F. Liddle, A. M. Smith, A. Mendrek, B. B. Forster, and R. D. Hare (1999). Neural pathways involved in the processing of concrete and abstract nouns. *Human Brain Mapping 7*, 225–233.

Malandrakis, N., A. Potamianos, E. Iosif, and S. Narayanan (2011). Kernel models for affective lexicon creation. In *Proc. Interspeech*, pp. 2977–2980.

Meng, H. and K.-C. Siu (2002). Semi-automatic acquisition of semantic structures for understanding domain-specific natural language queries. *IEEE Transactions on Knowledge and Data Engineering 14*(1), 172–181.

Mihalcea, R. and D. Radev (2011). *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge University Press.

Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography 3*(4), 235–312.

Miller, G. and W. Charles (1998). Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*(1), 1–28.

Noppeney, U. and C. J. Price (2004). Retrieval of abstract semantics. *NeuroImage 22*, 164–170.

Pado, S. and M. Lapata (2007). Dependency-based construction of semantic space models. *Computational Linguistics 33*(2), 161–199.

Paivio, A. (1971). *Imagery and Verbal Processes*. New York, Holt, Rinehart and Winston.

Papagno, C., R. Capasso, and G. Miceli (2009). Reversed concreteness effect for nouns in a subject with semantic dementia. *Neuropsychologia 47*(4), 1138–1148.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxanomy. In *Proc. of International Joint Conference for Artificial Intelligence*, pp. 448–453.

Rubenstein, H. and J. B. Goodenough (1965). Contextual correlates of synonymy. *Communications of the ACM 8*(10), 627–633.

Turney, P. (2006). Similarity of semantic relations. *Computational Linguistics 32*(3), 379–416.

Véronis, J. (2004). Hyperlex: Lexical cartography for information retrieval. *Computer Speech and Language 18*(3), 223–252.

Widdows, D. and B. Dorow (2002). A graph model for unsupervised lexical acquisition. In *Proc. of the 19th International Conference on Computational Linguistics*, pp. 1093–1099.

Finite State Temporality and Context-Free Languages

Derek Kelleher
Trinity College Dublin
kellehdt@tcd.ie

Carl Vogel
Trinity College Dublin
vogel@tcd.ie

Abstract

In the finite-state temporality approach, events in natural language semantics have been characterized in regular languages, with strings representing sequences of temporal observations. We extend this approach to natural language constructions which are not regular. Context-free constructions are detailed and discussed. Superposition, the key operator in the finite-state temporality approach is investigated for context-free languages. The set of context-free languages is found to not be closed under superposition. However, as with intersection, the superposition of a context-free language and a regular language results in a context-free language. Previous work on subsumption and entailment is inapplicable to context-free languages, due to the undecidability of the subset relation for context-free languages.

1 Introduction

In recent years, events have been encoded as strings of a regular language, where a symbol in the language represents a set of predicate logic formulae that hold at a particular temporal instant, and the order of the symbols is associated with temporal order. Temporal granularity is revealed by the "superposition" of languages, essentially forming a more specific representation by collecting together formulae that hold at the same time. The representation is supplemented by notions of subsumption and entailment, allowing comparisons of information content, logical soundness, and completeness.

However, some natural language constructions concerning events are difficult to represent in this regular framework. These constructions suggest a relationship between the frequency of events, similar to the dependency of symbol frequencies on other symbol frequencies found in many context-free languages ($L = a^n b^n$). Together with their increase in complexity over regular langauges, this makes context-free languages a natural area of interest in this field.

As an example, take the expression "A as often as B", where A and B can be thought of as events. This construction implies a frequency relationship between the occurrence of two events, where A occurs at least as many times as B, possibly more. These events do not have to occur in an ordered sequence, where a sequence of As are followed by a sequence of Bs, As and Bs can occur in any order as long as the overall frequncy relationship is maintained. To be accurate, we must also allow for "instants" of time that separate any occurrence of the events we are interested in. Representing event A as the symbol a, event B as the symbol b, and an instant of time in which events may be occurring, but which are not relevant to our analysis by $\Box$, we get strings of the form $a\Box^*b$, $a\Box^*b\Box^*b\Box^*a$, $b\Box^*b\Box^*a\Box^*a\Box^*a$ etc. These strings form a context-free language.

Moens and Steedman (1988) highlight the complicated nature of the phrase "when". They suggest that "When A, B" implies not a strictly temporal relationship, but a causal one, making it a prime candidate for representation by a context-free language. Note that "When I swear, I put money into the swear jar" implies that if I swear twice, I put money into the jar twice, but not necessarily in a particular temporal order. I may swear twice during the day, and have to wait until I get home to put money in the jar.

More formal (and seemingly less natural) constructions such as "an equal number of times as", while rare, do have a place in more formal literature such as legal documents. This particular construction appears in locations as varied as “The Federal Code of the United States of America” and the bye-laws of the town of New Canaan,CT: “he shall choose alternates in rotation so that the alternates chosen by the Chairman shall be seated as nearly an equal number of times as is possible”.[1]

Our analysis is restriced to the case of there being a frequency relationship between two types of events. It should be noted that the addition of a third event would lead to strings characterised by languages that are not context-free, similar to the difference between $a^n b^n$ and $a^n b^n c^n$. While these constructions tend to seem less natural (“I cried as often as I laughed, and I laughed as often as I sang”), they cannot be discounted.

The above linguistic data, while by no means exhaustive, provides a steady base from which to explore context-free languages in a finite-state temporality framework. The ubiquity of the phrases “when” or “whenever” highlights the need for this extension, while their causal nature, as opposed to temporal nature, suggests further ontological applications.

2 Background

Envisaging events as a sequence of “snapshots”, Fernando (2004) has encoded event-types as regular languages, made up of symbols representing sets of “fluents”(Φ), similar to those found in McCarthy and Hayes (1969). As well as representing event types, a regular language can represent sequences of temporal observations. The diagram below represents these two concepts:

$$L = \boxed{\sim swim(john, x)} \quad \boxed{swim(john, x)}^{+} \quad \boxed{swim(john, m)}$$

$$L' = \qquad\qquad \boxed{mile(m)}^{*}$$

The “superposition” of two langauges is the componentwise union of their strings:

$$L\&L' = \bigcup_{k\geq 1} \{(\alpha_1 \cup \alpha_1^{'}) \ldots (\alpha_k \cup \alpha_k^{'}) \mid \alpha_1 \ldots \alpha_k \in L \text{ and } \alpha_1^{'} \ldots \alpha_k^{'} \in L'\}$$

Intuitively, snapshots taken at the same temporal instant are merged, forming a larger picture of the world at that time:

$$\boxed{\sim swim(john, x) mile(m)}\boxed{swim(john, x) mile(m)}^{+}\boxed{swim(john, m) mile(m)}$$

The set of regular languages is closed under superposition ensuring that the superposition operation does not take us to a higher level of complexity(Fernando (2003)). Superposition allows us to define a reflexive, transitive relation (a pre-order) associated with the concept of subsumption. To preserve reflexivity subsumption $\trianglerighteq$ is defined by:

$$L \trianglerighteq L' \quad \text{iff} \quad L \subseteq L'$$

Subsuption can be thought of as relating to “information content”. A language that subsumes another is more specific than that language. It contains all the information of the other language, and more.

3 Superposition and Context-Free Languages

Superposition, as the central operation in the finite-state temporality framework, must be re-examined in light of our inclusion of context-free languages. The key question is whether the result of superposing a context-free language with either a regular language or another context-free language, is itself regular, context-free, or otherwise.

[1] http://ecode360.com/9045062 - accessed on 30/11/12.

Proposition 1 *The set of context-free languages is not closed under superposition.*

Proof(by counter-example): Let the set $\{\phi\}$ be represented by the symbol $\boxed{\phi}$, and the set $\{\psi\}$ be represented by the symbol $\boxed{\psi}$. The language $L_1 = \boxed{\phi}^n\boxed{\psi}^n$ is context-free, as is the language $L_2 = \boxed{\phi}^m\boxed{\psi}^{2m}$. L_1 is given by the grammar:

$S \rightarrow \boxed{\phi}S\boxed{\psi}$
$S \rightarrow e$

and L_2 by the grammar:

$S \rightarrow \boxed{\phi}S\boxed{\psi}\boxed{\psi}$
$S \rightarrow e$

The superposition of these two languages will contain strings consisting of three possible symbols: $\{\phi\} \cup \{\phi\} = \{\phi\}$ represented as $\boxed{\phi}$, $\{\phi\} \cup \{\psi\} = \{\phi, \psi\}$ represented as $\boxed{\phi\psi}$, and $\{\psi\} \cup \{\psi\} = \{\psi\}$ represented as $\boxed{\psi}$.

Strings in the language L_1 have length 2n, and strings in the language L_2 have length 3m. Strings can only be superposed if they have equal length, therefore only strings of length 6r from both languages can be superposed, resulting in strings of the same length. Strings in L_1 will consist of 3r $\boxed{\phi}$s followed by 3r $\boxed{\psi}$s, and strings in L_2 will consist of 2r $\boxed{\phi}$s followed by 4r $\boxed{\psi}$s. The superposition of these two strings will consist of 2r $\boxed{\phi}$s superposed with $\boxed{\phi}$s, r $\boxed{\phi}$s superposed with $\boxed{\psi}$s, and 3r $\boxed{\psi}$s superposed with $\boxed{\psi}$s, resulting in strings of the form $\boxed{\phi}^{2r}\boxed{\phi\psi}^{r}\boxed{\psi}^{3r}$. Introducing a homomorphism from $\boxed{\phi}$ to 'a', from $\boxed{\phi\psi}$ to 'b', and from $\boxed{\psi}$ to 'c', we have an equivalent language $a^{2r}b^rc^{3r}$.

If this language were context-free, given that it is infinte, there would be some constant K such that any string longer than K would be representable as a string uvxyz such that v and y are not empty and are pumpable. If we choose the string $a^{2K}b^Kc^{3K}$ as a string longer than K, we should be able to factorize it in this manner. If we chose v to have both as and bs or both bs and cs, then upon pumping it, we would have bs before as or cs before bs, which would result in a string not in our language. The same considerations apply to choosing y. Therefore v and y must each contain only as, or only bs, or only cs. Pumping v and y would therefore increase the number of one or two of the symbols but not all three, thereby losing the frequency relationship between the three symbols. The language cannot be context-free.□

Proposition 2 *The superposition of a context-free language with a regular language is context-free.*

Proof: Given L_1, a context-free language, and L_2, a regular language, let $P = \langle Q_P, \Sigma, \Gamma, \Delta_P, q_{P0}, F_P\rangle$ be a pushdown-automaton accepting L_1 and $A = \langle Q_A, \Sigma, \delta_A, q_{A0}, F_A\rangle$ be a finite-state-automoton accepting L_2. Δ_P is the set of transitions of the form $(q_i, a, A) \rightarrow (q_j, \gamma)$ interpreted as: when in state q_i , with input symbol a, and symbol A at the top of the stack, go to state q_j and replace A by the string γ, and δ_A is the set of transitions of the form $(q_i, a) \rightarrow (q_j)$ interpreted as: when in state q_i with input symbol a, go to state q_j. We form a pushdown automaton $R = \langle Q_P \times Q_A, \Sigma, \Gamma, \Delta_{P\times A}, (q_{P0}, q_{A0}), F_P \times F_A\rangle$, with transitions $\Delta_{P\times A}$ constructed as follows:

1. If Δ_P contains a rule of the form $(q_0, e, e) \rightarrow (q_1, S)$, then $\Delta_{P\times A}$ contains a rule of the form $((q_0, q_0), e, e) \rightarrow ((q_1, q_0), S)$.

2. If Δ_P contains a rule of the form $(q_1, e, A) \rightarrow (q_1, \gamma)$, then $\Delta_{P\times A}$ contains rules of the form $((q_1, q_x), e, A) \rightarrow ((q_1, q_x), \gamma)$ for every $q_x \in Q_A$.

3. If Δ_P contains a rule of the form $(q_1, a, a) \rightarrow (q_1, e)$, then $\Delta_{P\times A}$ contains rules of the form $((q_1, q_x), a \cup b, a) \rightarrow ((q_1, q_y), e)$ if and only if there is a transition $(q_x, b){\rightarrow}(q_y)$ in δ_A.

The new transitions are akin to running the PDA and FSA in tandem, where a state (q_x, q_y), while strictly a state of R, can be thought to represent the simultaneous states of P and A. A rule of type 1 and rules of type 2 perform the same stack operations as the PDA they were derived from. Therefore, R can produce on its stack the same set of strings that P produces on its stack. No input symbol is being read while these stack operations are performed, therefore R should remain in state (q_x, q_y). Rules of type 3 ensure that if R is in a state (q_1, q_y) with an input symbol $a \cup b$, and encounters the terminal symbol a on its stack, along with there being a transition in A from q_y to q_z on input b, then R will transition to state (q_1, q_z), and delete a from its stack. These are exactly the states that P and A would seperately be in upon reading input a and b respectively. Thus, if P reads a string $a_1 \dots a_n$ and is in a final state with an empty stack (i.e. P accepts this string), and A reads a string $b_1 \dots b_n$ and is in a final state (i.e. A accepts this string), then R will be in a final state upon reading the superposition of these two strings. If P accepts a language L_1 and A accepts a language L_2, then R will accept $L = L_1 \& L_2$.□

If we superpose the context-free language that represents "I laughed as often as I cried" with the regular langauge that represents "an hour" to get a language representing "In an hour, I cried as often as I laughed", this language, as the superposition of a context-free langauge and a regular language, will be context-free.

4 Final Remarks

Further work will involve investigating how the concepts of subsumption and entailment can be related to context-free languages. In this framework, entailment is defined in terms of subsumption, which is defined in terms of the subset relation (Fernando and Nairn (2005)). However, according to Hopcroft et al. (1979), the problem of whether a context-free language is a subset of another context-free language is undecidable. If the subset relation cannot be calculated for context-free languages, subsumption and entailment relations break down.

One possible avenue of exploration is the making of regular approximations of context-free languages (Mohri et al. (2001)). This would preserve the subsumption and entailment relations, but at a possible cost to accurately representing the context-free construction, possibly losing the exact relationship between the frequencey of two symbols.

References

Fernando, T. (2003). Finite-state descriptions for temporal semantics. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5).*

Fernando, T. (2004). A finite-state approach to events in natural language semantics. *Journal of Logic and Computation 14*(1), 79–92.

Fernando, T. and R. Nairn (2005). Entailments in finite-state temporality. In *Proc. 6th International Workshop on Computational Semantics*, pp. 128–138.

Hopcroft, J., R. Motwani, and J. Ullman (1979). *Introduction to automata theory, languages, and computation*, Volume 2. Addison-wesley Reading, MA.

McCarthy, J. and P. Hayes (1969). Some philosophical problems from the standpoint of artificial intelligence. In M. Meltzer and D. Michie (Eds.), *Machine Intelligence 4*. Edinburgh University Press.

Moens, M. and M. Steedman (1988). Temporal ontology and temporal reference. *Computational Linguistics 14*(2), 15–28.

Mohri, M., M. Nederhof, et al. (2001). Regular approximation of context-free grammars through transformation. *Robustness in language and speech technology 17*, 153–163.

Logic Programs vs. First-Order Formulas in Textual Inference

Yuliya Lierler
University of Nebraska at Omaha
ylierler@unomaha.edu

Vladimir Lifschitz
University of Texas at Austin
vl@cs.utexas.edu

Abstract

In the problem of recognizing textual entailment, the goal is to decide, given a text and a hypothesis expressed in a natural language, whether a human reasoner would call the hypothesis a consequence of the text. One approach to this problem is to use a first-order reasoning tool to check whether the hypothesis can be derived from the text conjoined with relevant background knowledge, after expressing all of them by first-order formulas. Another possibility is to express the hypothesis, the text, and the background knowledge in a logic programming language, and use a logic programming system. We discuss the relation of these methods to each other and to the class of effectively propositional reasoning problems. This leads us to general conclusions regarding the relationship between classical logic and answer set programming as knowledge representation formalisms.

1 Introduction

In the problem of recognizing textual entailment, the goal is to decide, given a text T and a hypothesis H expressed in a natural language, whether a human reasoner would call H a consequence of T. The following example is No. 115 in the collection of problems proposed as the Second PASCAL Recognizing Textual Entailment Challenge Bar-Haim et al. (2006):

> T: The World Bank has also been criticized for its role in financing projects that have been detrimental to human rights and the natural environment.
>
> H: The World Bank is criticized for its activities.
>
> *Expected answer*: Yes.

Recognizing textual entailment is a special case of a more general and practically important problem, textual query answering.

To recognize the fact that H is "entailed" by T, we often need to use some background commonsense knowledge. For instance, in the example above it is essential that financing is an activity.

The approach to recognizing textual entailment employed in Bos and Markert (2005) and implemented in the system Nutcracker[1] can be summarized as follows:

(i) T and H are represented first by discourse representation structures Kamp and Reyle (1993) and then by first-order formulas,

(ii) potentially relevant background knowledge is identified and expressed by a first-order formula *BK*,

(iii) an automated reasoning system is used to check whether the implication

$$T \wedge BK \rightarrow H \tag{1}$$

is logically valid.

[1] http://www.cogsci.ed.ac.uk/~jbos/RTE/.

Related work is described in Akhmatova (2005); Fowler et al. (2005).

The approach to the problem proposed in Baral et al. (2005); Tari and Baral (2005); Nouioua and Nicolas (2006) is similar, except that it relies on logic programs as the representation language instead of first-order formulas, and on logic programming systems as computational tools instead of first-order reasoners. The following example comes from the introduction to Baral et al. (2005):

> T: In Paris, on March 15th, John packed his laptop in the carry-on luggage and took a plane to Baghdad.
>
> H: His laptop was in Baghdad on March 16th.
>
> *Expected answer*: Yes.

Here again some background commonsense knowledge is needed to recognize that *yes* is the correct answer: a trip from Paris to Baghdad by air does not normally take more than a day; a person and his carry-on luggage are normally in the same city. Baral *et al.* represent the text T by a set of rules,[2] along with background knowledge *BK*; H is represented by a ground atom. Then an answer set solver and a constraint logic programming system are used to establish the fact that H is entailed by logic program $T \cup \mathit{BK}$.

Each of the two knowledge representation languages—first-order formulas and logic programs—has its own advantages. A first-order formula may have a complex, nested form; this is essential because discourse representation structures are often deeply nested. On the other hand, the semantics of logic programs is nonmonotonic; this is crucial when background commonsense knowledge is expressed by defaults (note the word "normally" in the examples above).

In this paper we argue, however, that these two versions of the logic approach to textual entailment have more in common than meets the eye. A large part of the work done by Bos and Markert can be understood in terms of the logic programming methodology advocated by Baral *et al.* Many textual entailment problems used to test the Nutcracker system can be solved by the answer set solver DLV[3] instead of the first-order theorem prover VAMPIRE[4] and the model builder PARADOX[5] that were actually employed in Nutcracker experiments.

The first-order reasoning problems that can be naturally expressed by logic programs have a distinctive syntactic feature: they belong to the "effectively propositional," or "near-propositional" formulas Schulz (2002). The relationship between effectively propositional reasoning (EPR) and answer set programming (ASP) Lifschitz (1999); Marek and Truszczyński (1999); Niemelä (1999) is one of the topics discussed in this paper. This will bring us, at the end of the paper, to some general conclusions regarding the relationship between classical logic and answer set programming as knowledge representation formalisms.

2 Representing EPR Formulas by Logic Programs

We consider here first-order formulas that may contain equality and object constants, but not function constants of arity > 0. An *EPR formula* is the universal closure of a quantifier-free formula in conjunctive normal form. We will show how to turn any EPR formula F into a logic program $\pi(F)$ such that $\pi(F)$ has a stable model iff F is satisfiable.

In the definition of π we assume that every clause in F is written as an implication with a conjunction of atoms (possibly empty) in the antecedent, and a disjunction of atoms (possibly empty) in the consequent:

$$A_1 \wedge \cdots \wedge A_m \rightarrow A_{m+1} \vee \cdots \vee A_n. \tag{2}$$

[2] This is done manually; automating the translation is mentioned in the paper as future work.

[3] `http://www.dbai.tuwien.ac.at/proj/dlv/.`

[4] `http://en.wikipedia.org/wiki/Vampire_theorem_prover.`

[5] `http://www.math.chalmers.se/~koen/paradox/.`

Besides the predicate constants occurring in F, the program $\pi(F)$ will contain two new predicate constants: the unary constant u (for "universe") and the binary constant eq (for "equals"). For any atomic formula A, by A^{eq} we denote $eq(t_1, t_2)$ if A is an equality $t_1 = t_2$, and A otherwise.

Program $\pi(F)$ consists of

(i) the facts $u(c)$ for all object constants c occurring in F;

(ii) the rules

$$\begin{array}{l} eq(X,X) \leftarrow u(X) \\ eq(Y,X) \leftarrow eq(X,Y) \\ eq(X,Z) \leftarrow eq(X,Y), eq(Y,Z); \end{array}$$

(iii) the rules

$$p(Y_1, \dots, Y_k) \leftarrow p(X_1, \dots, X_k), eq(X_1, Y_1), \dots, eq(X_k, Y_k)$$

for all predicate constants p occurring in F;

(iv) the disjunctive rules

$$A^{eq}_{m+1}; \dots ; A^{eq}_n \leftarrow A^{eq}_1, \dots, A^{eq}_m, u(X_1), \dots, u(X_k)$$

corresponding to the conjunctive terms (2) of F, where $X_1, \dots, X_k$ are the variables that occur in the consequent $A_{m+1} \vee \dots \vee A_n$ of (2) but do not occur in its antecedent $A_1 \wedge \dots \wedge A_m$.

If, for instance, F is

$$\forall X (p(a) \wedge p(b) \wedge \neg p(c) \wedge (p(X) \rightarrow X = a)) \tag{3}$$

then $\pi(F)$ consists of the rules

$$\begin{array}{l} u(a) \leftarrow \qquad u(b) \leftarrow \qquad u(c) \leftarrow \\ eq(X,X) \leftarrow u(X) \\ eq(Y,X) \leftarrow eq(X,Y) \\ eq(X,Z) \leftarrow eq(X,Y), eq(Y,Z) \\ p(Y) \leftarrow p(X), eq(X,Y) \\ p(a) \\ p(b) \\ \leftarrow p(c) \\ eq(X,a) \leftarrow p(X). \end{array} \tag{4}$$

As usual in answer set programming, by a model (stable model) Gelfond and Lifschitz (1988, 1991) of $\pi(F)$ we understand a model (stable model) of the corresponding ground program. It is clear that $\pi(F)$ is a (possibly disjunctive) program without negation. Its stable models are simply minimal models. Each rule of $\pi(F)$ is safe—every variable occurring in its head occurs also in its body. The stable models of this program can be generated by the answer set solver DLV.

Theorem *For any EPR formula F, the following conditions are equivalent:*

(i) F is satisfiable,

(ii) $\pi(F)$ has a model,

(iii) $\pi(F)$ has a stable model.

If F does not contain equality then this assertion remains valid if all rules containing eq are dropped from $\pi(F)$.

For instance, formula (3) is satisfiable; accordingly, program (4) has a stable model:

$$\{u(a),\ u(b),\ u(c),\ p(a),\ p(b),\ eq(a,a),\ eq(b,b),\ eq(c,c),\ eq(a,b),\ eq(b,a)\}.$$

Proof We begin by proving the second claim, that is, assume that F does not contain equality and that the rules containing *eq* are dropped from the program $\pi(F)$. *From (i) to (ii).* F is the universal closure of a quantifier-free formula, and it is satisfiable. Consequently F has an Herbrand model. By adding to this model the atoms $u(c)$ for all object constants c we get a model of $\pi(F)$. *From (ii) to (iii).* The result of grounding $\pi(F)$ is a finite program without negation; since it has a model, it has a minimal model. *From (iii) to (i).* By removing the atoms $u(c)$ from a model of $\pi(F)$ we get an Herbrand model of F. Consider now the general case, when F may contain equality. Let F^* be the formula obtained from F by

- replacing each equality $t_1 = t_2$ with $eq^*(t_1, t_2)$, where eq^* is a new binary predicate constant;
- conjoining the result with the universal closures of the formulas

$$\begin{array}{l} eq^*(X,X), \\ eq^*(X,Y) \rightarrow eq^*(Y,X), \\ eq^*(X,Y) \land eq^*(Y,Z) \rightarrow eq^*(X,Z) \end{array}$$

  and

$$p(X_1,\dots,X_k) \land eq^*(X_1,Y_1) \land \cdots \land eq^*(X_k,Y_k) \rightarrow p(Y_1,\dots,Y_k)$$

  for all predicate constants p occurring in F;
- converting the result to prenex form.

It is clear that F^* is an EPR formula that does not contain equality, and that it is satisfiable iff F is satisfiable. The rules of $\pi(F^*)$ that do not contain *eq* can be obtained from the rules of $\pi(F)$ by replacing each occurrence of *eq* with eq^*. Consequently $\pi(F^*)$ has a model (stable model) iff $\pi(F)$ has a model (stable model). It remains to apply the part of the theorem proved above to F^*.

This proof shows that, in the case when F does not contain equality, the models of $\pi(F)$ are essentially identical to the Herbrand models of F.

From the theorem we see that testing an EPR formula for satisfiability can be accomplished using an answer set solver. In the next section we investigate the applicability of this idea to the problem of recognizing textual entailment.

3 EPR Formulas in Experiments with Textual Entailment

Recall that Bos and Markert [2005] recognize textual entailment by determining whether implication (1) is logically valid.[6] In many cases, the negation

$$T \land BK \land \neg H \tag{5}$$

of formula (1) can be converted to a prenex form with all existential quantifiers in front of the universal quantifiers ("$\exists\forall$-prenex form"). Then the sentence F, obtained from this prenex form by Skolemization and then converting the quantifier-free part to conjunctive normal form, is an EPR formula. It is clear that (1) is logically valid iff F is unsatisfiable.

The possibility of converting (5) to $\exists\forall$-prenex form is essential because it guarantees that no function constants of arity > 0 are introduced in the process of Skolemization. It is clear that conjunction (5) can be written in $\exists\forall$-prenex form if every conjunctive term can be written in this form.

[6] Like other existing systems for recognizing textual entailment, Nutcracker is not completely reliable. Generally, formulas H and T only approximate the meanings of RTE sentence pairs. For instance, Nutcracker currently ignores the semantics of plurals, tense, and aspect. Also, formula *BK* may not adequately represent all relevant background knowledge.

How restrictive is this requirement? The website shown in Footnote 1 gives the first-order formulas corresponding to 799 textual entailment problems from the second PASCAL collection that were produced by Nutcracker. In 711 cases (89%), both T and $\neg H$ can be converted to $\exists\forall$-prenex form. The conjunctive terms of *BK* come from two sources: most are automatically extracted from WordNet, the others are hand-coded. All conjunctive terms of the first kind are universal sentences, so that each of them is $\exists\forall$-prenex. Among the hand-coded parts of *BK* that Bos and Markert chose to include we found several exceptions, but they are never essential: dropping the "difficult" hand-coded part of *BK* from any of the 711 implications (1) that we have studied never makes a logically valid formula invalid.

The model builder PARADOX terminates, in principle, on any EPR formula; it generates a model if the formula is satisfiable, and reports that it is unsatisfiable otherwise. For this reason, in each of the 711 cases (with the inessential "difficult" terms dropped) Bos and Markert would be able to test the formula for satisfiability by running PARADOX alone, without invoking also the theorem prover VAMPIRE.

In these 711 cases we could also test the logical validity of (1) by running the answer set solver DLV, as described in the previous section. Interestingly, the runtime of DLV was smaller than the runtime of the model builder PARADOX in most cases, even though DLV did some redundant work: whenever the program $\pi(F)$ has a model, it computed one of its *minimal* models.[7] We should add that all these runtimes are pretty small, usually less than a second.

The main computational difference between model builders, such as PARADOX, and answer set solvers, such as DLV, is that model builders typically start by looking for a finite model with a small universe (say, a singleton); if there is no such model then search is extended to larger universes. If the input is an EPR formula containing N object constants, then answer set solvers ground the given program with the universe of size N, which means essentially that they only look for a model of cardinality N. Good answer set solvers, such as DLV, perform grounding "intelligently" and sometimes introduce auxiliary predicates that make the size of the grounded program smaller.

It is interesting also that among the 711 EPR formulas from Nutcracker experiments that we have investigated, 514 are Horn—all their conjunctive terms (2) are definite ($n = m+1$) or negative ($n = m$). If F is a Horn EPR formula then the program $\pi(F)$ is nondisjunctive; since there is no negation in this program, it can have at most one stable model. Note that deciding whether a ground nondisjunctive program without negation has a model can be done in linear time.

4 Conclusion

The properties of the translation π established in this paper suggest that EPR reasoning can be described as the common part of classical first-order logic and logic programming under the stable model semantics: (Figure 1). In terms of expressiveness, the availability of formulas more complex than EPR is a strong side of classical logic; the availability of negation as failure is a strong side of declarative logic programming. Representations used in the existing work on the analysis of textual inference in terms of classical logic belong mostly to the common part of the two areas.

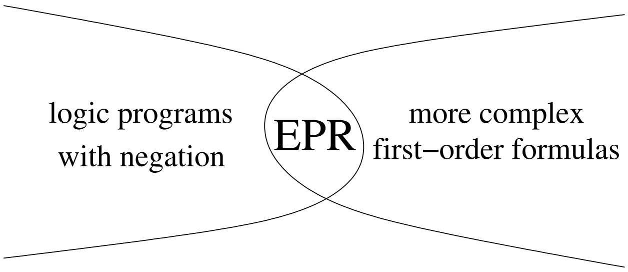


Figure 1: Logic programs vs. first-order formulas: a comparison

[7] System DLV has the option `-OM-` that disables testing for minimality, but in our experiments it did not have a noticeable effect on runtimes.

References

Akhmatova, E. (2005). Textual entailment resolution via atomic propositions. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Bar-Haim, R., I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor (2006). The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Baral, C., G. Gelfond, M. Gelfond, and R. Scherl (2005). Textual inference by combining multiple logic programming paradigms. In *AAAI Workshop on Inference for Textual Question Answering*.

Bos, J. and K. Markert (2005). Recognising textual entailment with logical inference. In *Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 628–635.

Fowler, A., B. Hauser, D. Hodges, I. Niles, A. Novischi, and J. Stephan (2005). Applying COGEX to recognize textual entailment. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Gelfond, M. and V. Lifschitz (1988). The stable model semantics for logic programming. In R. Kowalski and K. Bowen (Eds.), *Proceedings of International Logic Programming Conference and Symposium*, pp. 1070–1080. MIT Press.

Gelfond, M. and V. Lifschitz (1991). Classical negation in logic programs and disjunctive databases. *New Generation Computing 9*, 365–385.

Kamp, H. and U. Reyle (1993). *From discourse to logic*, Volume 1,2. Kluwer.

Lifschitz, V. (1999). Action languages, answer sets and planning. In *The Logic Programming Paradigm: a 25-Year Perspective*, pp. 357–373. Springer Verlag.

Marek, V. and M. Truszczyński (1999). Stable models and an alternative logic programming paradigm. In *The Logic Programming Paradigm: a 25-Year Perspective*, pp. 375–398. Springer Verlag.

Niemelä, I. (1999). Logic programs with stable model semantics as a constraint programming paradigm. *Annals of Mathematics and Artificial Intelligence 25*, 241–273.

Nouioua, F. and P. Nicolas (2006). Using answer set programming in an inference-based approach to natural language semantics. In *Proceedings of the Fifth Workshop on Inference in Computational Semantics (ICoS)*.

Schulz, S. (2002). A comparison of different techniques for grounding near-propositional CNF formulae. In *Proceedings of the 15th International FLAIRS Conference*, pp. 72–76.

Tari, L. and C. Baral (2005). Using AnsProlog with Link Grammar and WordNet for QA with deep reasoning. In *AAAI Workshop on Inference for Textual Question Answering*.

Using Network Approaches to Enhance the Analysis of Cross-Linguistic Polysemies

Johann-Mattis List
Philipps-University Marburg
mattis.list@uni-marburg.de

Anselm Terhalle
Heinrich Heine University Düsseldorf
terhalle@phil.hhu.de

Matthias Urban
Philipps-University Marburg
matthias.urban@uni-marburg.de

Abstract

Since long it has been noted that cross-linguistically recurring polysemies can serve as an indicator of conceptual relations, and quite a few approaches to model and analyze such data have been proposed in the recent past. Although – given the nature of the data – it seems natural to model and analyze it with the help of network techniques, there are only a few approaches which make explicit use of them. In this paper, we show how the strict application of weighted network models helps to get more out of cross-linguistic polysemies than would be possible using approaches that are only based on item-to-item comparison. For our study we use a large dataset consisting of 1252 semantic items translated into 195 different languages covering 44 different language families. By analyzing the community structure of the network reconstructed from the data, we find that a majority of the concepts (68%) can be separated into 104 large communities consisting of five and more nodes. These large communities almost exclusively constitute meaningful groupings of concepts into conceptual fields. They provide a valid starting point for deeper analyses of various topics in historical semantics, such as cognate detection, etymological analysis, and semantic reconstruction.

1 Introduction

What do “milk” and “udders” have to do with each other? Conceptually, they are closely related, since the former is the product and the content of the latter. Linguistically, they may even look the same, being referred to by identical word forms in many different languages, such as, e.g., by [nax] in Judeo-Tat (an Indo-European language), by [ukun] in Oroqen (an Altaic language), or by [mis] in Miao (a Hmong-Mien language, all data taken from Key and Comrie 2007). Historically, the conceptual relation between “milk” and “udders” may show up in the form of semantic shifts where a word which was formerly used to express one of the concepts in a given language is henceforth used to express the other one. Thus, in Standard Chinese, the word [niou^{35}nai^{214}] “milk” is a compound of [niou^{35}] “cow” and [nai^{214}] “milk” which originally meant “udder” (as well as “breast”).[1]

The situation in which a set of conceptually related meanings is expressed by the same form in a given language is known as *polysemy*. In this paper, we show how an analysis of a weighted network reconstructed from a large database of cross-linguistic polysemies can help to shed light on common conceptual associations in the world’s languages. We find that the reconstructed network has a strong *community structure*: it can be easily separated into communities, i.e. groups of nodes that share more connections with each other than with nodes outside the group. We detected 104 large communities consisting of five and more nodes. These communities cover 879 out of 1286 concepts (68%) and almost exclusively constitute meaningful groupings of concepts into conceptual fields, thus providing a valid starting point for further analyses in historical linguistics.

[1] Superscript numbers indicate tones.

2 Polysemy and Conceptual Relations in Synchrony and Diachrony

The term *polysemy* was first used by Bréal (1897, 154), who explicitly introduced the notion as a direct consequence of semantic change. Indeed, recent approaches to semantic change (e.g. Traugott and Dasher 2002) emphasize that the development of an initially secondary polysemous reading is the first stage of a complete semantic change to take place. Given that for semantic change a certain 'association [...] between the old meaning and the new [...] might be regarded as a necessary condition' (Ullmann, 1972, 211), the same kind of "association" can also be assumed to hold for polysemy. There are recent approaches in historical semantics that explicitly start from polysemous lexical items to extract information about what kinds of concepts are associated and what kinds of semantic change seem to be plausible (Croft et al., 2009). This procedure has the great advantage of providing 'an important antidote to the unbridled imagination in postulating etymologies' (Evans and Wilkins, 2000, 550), where intuitive assessments regarding plausibility as far as semantics is concerned are still frequent. Our approach operates in a similar vein.

3 Modeling Cross-Linguistic Polysemies as Weighted Networks

The idea to model polysemies as networks itself is not new. It was already underlying Haspelmath's (2003) *semantic map* approach which is used as a heuristic tool to analyze grammatical categories in linguistic typology. François (2008) applied this approach to the lexical domain, followed by further work by Croft et al. (2009), Perrin (2010), Cysouw (2010a, 2010b), and Steiner et al. (2011), who also introduced a simplified procedure to retrieve putative polysemies from semantically aligned word lists. What is new in our approach, however, is the strict modeling of cross-linguistic polysemies as *weighted* networks that shall be briefly introduced in the following.

3.1 Reconstruction

Networks (or graphs) constitute a system representation tool which is used by many different disciplines (Newman, 2004). We make use of a weighted network model to display and analyze conceptual relations reconstructed on the basis of cross-linguistic polysemies. Our network has the following structure:

> Let C be a set of n concepts $c_1, \ldots, c_n$ whose linguistic representation we want to analyze on the basis of a set L of m different languages belonging to $o \leq m$ language families. Our network is an undirected weighted graph $G = (V, E, f)$, with $V = C$ and $E \subseteq \{e_{ij} := \{c_i, c_j\} \mid c_i, c_j \in C \text{ and } i \neq j\}$. f is a mapping from E into $\mathbb{N}$ with $f(e_{ij}) \leq o$ being the number of language families which use one word for both c_i and c_j, and $f(E) = \{f(e_{ij}) \mid e_{ij} \in E\}$.[2]

In less formal terms, we reconstruct a weighted network by representing all concepts in a given multilingual word list as nodes (vertices), and draw edges between all nodes that show up as polysemies in the data. The edge weights reflect the number of language families in which these polysemies are attested.

3.2 Analysis

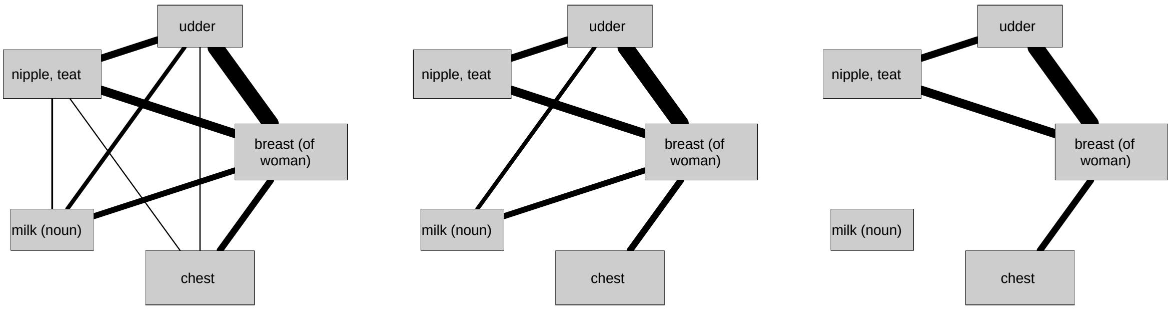
Statistical accounts on cross-linguistic polysemies retrieved from semantically aligned word lists make it possible to define the similarity between concepts on an item-to-item basis. Here, problems may arise from the fact that our approach cannot make an a priori distinction between polysemy and semantic vagueness on the one hand (Geeraerts, 1993), and polysemy and homophony on the other (compare François' 2008 notion of *colexification*). While, as Haspelmath (2003, 231) and François (2008, 169f) argue, the distinction between (true) polysemy and semantic vagueness does not matter from a cross-linguistic perspective (and we will make no attempt to distinguish the two), the failure to single out homophones can lead to wrong assessments regarding concept relations. In German, Dutch, and Yiddish, for example, the stems expressing the concepts "arm" and "poor" are identical in form. If one naively counted the number

[2] Many thanks to Daniel Schulzek for helpful comments on the maths.

of languages where the concepts are expressed by the same word, the link would appear to find much more support in the languages of the world than that between "udder" and "chest", which is only reflected in two languages in our data, namely in Aymara and Sirionó (see Supplemental Material), although in this latter case, the connection between the two concepts seems much more meaningful than in the former one.

A first way to solve this problem is to count occurrences of links between concepts not on the basis of languages but of language families, thus avoiding that they result from genetic inheritance. This would resolve the problem in favor of "udder" and "chest", since the link between "arm" and "poor" only occurs in the closely related Germanic languages, whereas Aymara and Sirionó belong to different language families. The problem can further be addressed by setting up a threshold of occurrences and to ignore links that exhibit less occurrences. Such an analysis is illustrated in Figure 1, where, starting from a given concept network, the threshold is successively increased and more and more edges are successively removed. While such an analysis surely yields the most reliable links, it has the drawback of ignoring many links that might reflect true – although less frequently attested – conceptual relations.

In order to solve the problem of separating the wheat from the chaff without losing too much wheat, the network perspective as opposed to the item-to-item perspective can be of great help. For the kind of networks we are dealing with in this study, an analysis of *community structure* seems to be specifically useful. Community structure refers to the property of many networks that do not consist of random collections of nodes with random connections between them but of 'distinct "communities" – groups of vertices within which the connections are dense but between which they are sparser' (Newman, 2004, 4). It is straightforward to assume that a network reflecting relations between concepts should exhibit some kind of community structure, given that it is often assumed that concepts can be grouped into specific conceptual fields. Analyzing the community structure of cross-linguistic polysemy networks should therefore give us some interesting insights into general concept relations.



(a) no cutoff (b) cutoff < 5 (c) cutoff < 10

Figure 1: Setting up a Threshold of Occurrences.

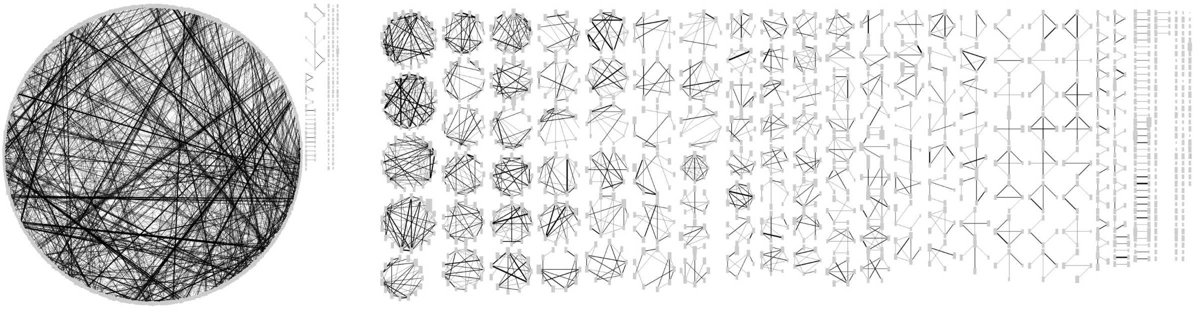
4 Application

4.1 Data

Our analysis is based on a large multilingual word list consisting of 1252 glosses ("concepts") translated into 195 different languages, covering 44 different language families (see Supplemental Material). The data was taken from three different sources, namely the Intercontinental Dictionary Series (IDS, Key and Comrie 2007, 133 languages), the World Loanword Database (WOLD, Haspelmath and Tadmor 2009, 30 languages), and a multilingual dictionary provided by the Logos Group (Logos, Logos Group 2008, 32 languages). The data from each of these sources was automatically cleaned and normalized with help of Python scripts. While the original sources of IDS and WOLD have a total of 1310 glosses, we selected only those glosses which were at least reflected in 100 languages. The structure of the input data is illustrated in Figure 2, with two instances of polysemies in Russian and German marked in bold font.

Key	Concept	Russian	German	...
1.1	world	mir, svet	Welt	...
1.21	earth, land	zemlja	**Erde**, Land	...
1.212	ground, soil	počva	**Erde**, Boden	...
1.420	tree	**derevo**	Baum	...
1.430	wood	**derevo**	Wald	...
...	...	...	...	...

Figure 2: Structure of the Data



(a) Complete Network

(b) Community Structure

Figure 3: Comparing the Network and its Community Structure.

4.2 Analysis

We created a weighted network from the data, using the number of language families in which a particular polysemy was reflected as edge weights. We further analyzed the community structure of the data with help of a weighted version (Newman, 2004) of the Girvan-Newman algorithm for community detection (Girvan and Newman, 2002).[3] This algorithm successively removes the edges with the highest *betweenness* from a given network. Edge betweenness is defined as 'the number of shortest paths between pairs of vertices that run along it' (Girvan and Newman, 2002, 7822). We followed Newman (2004) in using *modularity*, i.e. the 'fraction of edges that fall within communities minus the expected value of the same quantity if edges are assigned at random' (Newman, 2004, 6), as a criterion to find the best split for the analysis.

4.3 Results

In Figure 3 the original network (a) and the network's community structure (b) are contrasted. The original network consists of one very large connected component of 1141 nodes and only a spurious number of unconnected nodes. The analysis of the network with help of the Girvan-Newman algorithm yielded a total of 337 communities of which 104 are rather large, consisting of 5 and more nodes, and covering a majority of the concepts (879 out of 1289, 68%). Most of these large communities constitute meaningful groupings of concepts into conceptual fields. Community 5, for example, groups concepts that deal with the cover of bodies ("feather", "hair", "bark", etc.). Community 28 deals with learning ("study", "count", "try, attempt", etc.). And community 70 contains concepts related to transport vehicles ("canoe", "boat", and "carriage, wagon, cart", etc., see Supplemental Material).

Apart from the general question of which items are grouped together in one community, it is also interesting to investigate the internal structure of the communities more closely. Community 3, for example, consists of 18 items which all center around the concepts "tree" and "wood" (see the network representation using force-directed layout in Figure 4). Cross-linguistically, the conflation of "tree" and

[3] We are well aware of the fact that there are many other, supposedly better, algorithms for community detection. However, given that this is a pilot study, we decided to take a rather simple algorithm whose basic ideas are nicely described and easy to understand, thus limiting our presuppositions about conceptual structures to a minimum.

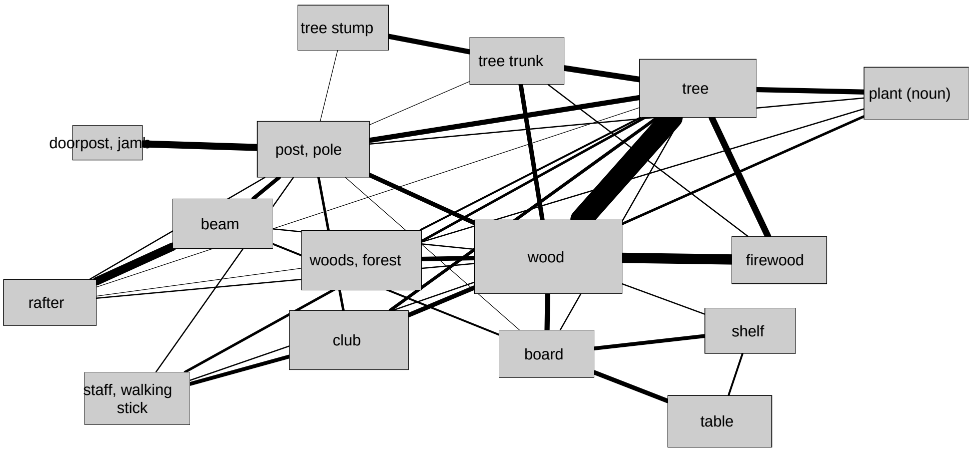


Figure 4: Force-Directed Representation of Cluster 3.

"wood" into one lexical form occurs very frequently, and it has been identified previously by several scholars (Hjelmslev, 1963; Witkowski et al., 1981). Our data is congruent with the traditional literature in this respect. However, in our network, these central concepts are connected with a substantial number of contiguously related ones. Thus, "wood" has something to do with "clubs", "staffs", "walking sticks", and the like, in that these all are artifacts "made from wood", and "trees" (and "wood") have also a contiguous relation with "forests" in so far as forests are, simply spoken, "agglomerations of trees". Unlike the well-documented association between "wood" and "tree", to our knowledge most of these further semantic connections have not yet been documented explicitly from a cross-linguistic point of view. Hence, as even this single example shows, our analysis contributes to enhancing our knowledge regarding cross-linguistically common conceptual associations. From the perspective of historical semantics, as Witkowski et al. (1981) argue, polysemies between "tree" and "wood" can be interpreted in diachronic terms, with terms for "wood" frequently giving rise to terms for "tree". But there is a diachronic correlate also for other concepts in this cluster. For example, the Proto-Indo-European root **dóru-* has a direct descendant in Russian *dérevo* "tree, wood". In other languages, however, reflexes are used to denote all kinds of weapons that were originally made from wood, such as Avestan *dāuru* "club" (but also "tree trunk"), or Modern Greek δόρυ "spear", which regularly meant "wood, tree" (but also "beam, pole") in Old Greek (Nussbaum 1968, 147, footnote).

5 Conclusion

Polysemies offer powerful evidence to study conceptual relations and semantic change. In this paper, we tried to show how the analysis of such data can be enhanced with the help of weighted network approaches. By applying them to a large dataset, we illustrated the potential of cross-linguistic polysemy data for semantic analyses. The analysis of the community structure of our network not only reproduces findings from the literature, but also reveals additional cross-linguistic regularities in the conceptual structures underlying semantic change.

Supplemental Material

The supplemental material accompanying this study contains the word lists of all 195 languages, upon which the reconstruction of the weighted network was based, and all essential results in form of text files, including the network, a detailed list and description of all inferred communities, and additional statistics regarding the languages.

References

Bréal, M. (1897). *Essai de sémantique*. Paris: Hachette.

Croft, W., C. Beckner, L. Sutton, T. Wilkins, J.and Bhattacharya, and D. Hruschka (2009). *Quantifying semantic shift for reconstructing language families*. Talk, held at the 83rd Annual Meeting of the Linguistic Society of America. PDF: `http://www.unm.edu/~wcroft/Papers/Polysemy-LSA-HO.pdf`.

Cysouw, M. (2010a). Drawing networks from recurrent polysemies. *Linguistic Discovery 8*(1), 281–285.

Cysouw, M. (2010b). Semantic maps as metrics on meaning. *Linguistic Discovery 8*(1), 70–95.

Evans, N. and D. Wilkins (2000). In the mind's ear: The semantic extensions of perception verbs in Australian languages. *Language 76*(3), 546–592.

François, A. (2008). Semantic maps and the typology of colexification: intertwining polysemous networks across languages. In M. Vanhove (Ed.), *From polysemy to semantic change*, pp. 163–215. Amsterdam: Benjamins.

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics 4*(3), 223–272.

Girvan, M. and M. E. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America 99*(12), 7821–7826.

Haspelmath, M. (2003). The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In M. Tomasello (Ed.), *The new psychology of language*, pp. 211–242. Mahwah, NJ: Lawrence Erlbaum.

Haspelmath, M. and U. Tadmor (2009). *World Loanword Database*. Munich: Max Planck Digital Library.

Hjelmslev, L. (1963). *Prolegomena to a theory of language*. Madison: University of Wisconsin Press.

Key, M. R. and B. Comrie (2007). *IDS – The Intercontinental Dictionary Series*. URL: `http://lingweb.eva.mpg.de/ids/`.

Logos Group (2008). *Logos Dictionary*. URL: `http://www.logosdictionary.org/index.php`.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E 70*(5), 056131.

Nussbaum, A. J. (1968). *Head and horn in Indo-European. The words for "horn," "head," and "hornet"*. Berlin and New York: de Gruyter.

Perrin, L.-M. (2010). Polysemous qualities and universal networks, invariance and diversity. *Linguistic Discovery 8*(1), 259–280.

Steiner, L., P. F. Stadler, and M. Cysouw (2011). A pipeline for computational historical linguistics. *Language Dynamics and Change 1*(1), 89–127.

Traugott, E. C. and R. B. Dasher (2002). *Regularity in semantic change*. Cambridge: Cambridge University Press.

Ullmann, S. (1972). *Semantics*. Blackwell.

Witkowski, S. R., C. H. Brown, and P. K. Chase (1981). Where do tree terms come from? *Man 16*(1), 1–14.

A corpus-based taxonomy of question responses

Paweł Łupkowski*
Institute of Psychology
Adam Mickiewicz University, Poznan
Pawel.Lupkowski@amu.edu.pl

Jonathan Ginzburg
UFR Études anglophones
Université Paris-Diderot (Paris 7)
yonatan.ginzburg@univ-paris-diderot.fr

Abstract

In this paper we consider the issue of answering a query with a query. Although these are common, with the exception of Clarification Requests, they have not been studied empirically. After briefly reviewing different theoretical approaches on this subject, we present a corpus study of query responses in the British National Corpus and develop a taxonomy for query responses. We sketch a formal analysis of the response categories in the framework of KoS.

1 Introduction

Responding to a query with a query is a common occurrence, representing on a rough estimate more than 20% of all responses to queries found in the British National Corpus.[1] Research on dialogue has long recognized the existence of such responses. However, with the exception of one of its subclasses—albeit a highly substantial one—the class of query responses has not been characterized empirically in previous work. The class that has been studied in some detail are Clarification Requests (CRs) (Rodriguez and Schlangen, 2004; Rieser and Moore, 2005). However, CRs can be triggered by any utterance, interrogative or otherwise. Researchers on the semantics and pragmatics of questions (see e.g. Carlson, 1983; Wiśniewski, 2003) have been aware for many years of the existence of one class of query responses—responses that express questions dependent on the question they respond to, as in (1a,b). This lead Carlson to propose (1c) as a sufficient condition for a query response, which can be formalized using (1d), assuming notions of resolvedness and aboutness (for which see e.g. Ginzburg and Sag, 2000).

(1) a. **A:** Who murdered Smith? **B:** Who was in town?

b. **A:** Who is going to win the race? **B:** Who is going to participate?

c. Who killed Smith depends on who was in town at the time.

d. q_2 can be used to respond to q_1 if q_1 $\mathsf{depends\ on}$ q_2.

e. q_1 $\mathsf{depends\ on}$ q_2 iff any proposition p such that p $\mathsf{Resolves}$ q_2, also satisfies p entails r such that r is About q_1.

Larsson (2002) and Asher and Lascarides (2003) argue that the proper characterization of question responses is pragmatically based. Asher and Lascarides (2003) propose to characterize non-CR query responses by means of the rhetorical relation *question elaboration* (Q-Elab) with stress on the plan-oriented relation between the initial question and the question expressed by the response. Q-Elab might be informally summarized as follows:

(2) If Q-Elab(α, β) holds between an utterance α uttered by A, where g is a goal associated by convention with utterances of the type α, and the question β uttered by B, then any answer to β must elaborate a plan to achieve g.

* The author was supported by the Iuventus Plus grant (IP2011-031-771).

[1] In the spoken part of the BNC, using SCoRE (Purver, 2001), we found 11312 ?/? cross-turn sequences, whereas 41041 ?/. cross-turn sequences, so the ?/? pairs constitute 21.6%. (For the SCoRE syntax see `http://www.dcs.qmul.ac.uk/imc/ds/score/help.html`.)

Table 1: Tags used to annotate question—question-response sample

Tag	Question-response category
`CR`	clarification requests
`DP`	dependent questions
`FORM`	questions considering the way of answering $q1$
`MOTIV`	questions about the underlying motivations behind asking $q1$
`NO ANSW`	questions aimed at avoiding answering $q1$
`QA`	questions providing an answer to $q1$
`IGNORE`	questions ignoring $q1$
`IND`	questions with a presupposed answer

The relation of Q-Elab, motivated by interaction in cooperative settings, is vulnerable to examples such as those in (3). (3a) has one understanding that might be characterized using dependence (*What I like depends on what* YOU *like*), but can also be used simply as a coherent retort. (3b) could possibly be used in political debate without it necessarily involving an attempt to discover an answer to the first question asked.

(3) a. **A:** What do you like? **B:** What do you like?

b. **A:** What is Sarkozy going to do about it? **B:** What is Papandreou?

In order to better understand the nature of question responses, we ran a corpus study on the British National Corpus (BNC). The results we obtained show that, apart from CRs, dependent questions are indeed by far the largest class of question responses. However, they reveal also the existence of a number of response categories, characterizable neither as dependent questions nor as plan supporting responses. They include (a) a class akin to what Conversation Analysts refer to as *counters* (Schegloff, 2007)—responses that attempt to foist on the conversation a distinct issue from the current discourse topic and (b) responses that ignore the current topic but address the situation it concerns.

Attaining completeness in characterizing the response space of a query is of fundamental importance for dialogue management and the design of user interfaces. Beyond that general goal, a better understanding of e.g. *counters* and *situation–relevant responses*, which we believe are rare in task–oriented dialogue, is important for adversarial interaction (courtroom, interrogation, argumentation, certain games). Characterizing their coherence is challenging for all approaches that ground dialogue on cooperativity principles (e.g. Asher and Lascarides, 2003; Roberts, 2011).

The rest of the paper is structured as follows: in section 2 we present the taxonomy underlying our corpus study; section 3 describes the results; in section 4 we sketch a formal analysis of one of the response categories in the framework of KoS (Ginzburg and Fernández, 2010). We conclude with a summary and future work.

2 A corpus-based taxonomy of answering by means of questions

The study sample The taxonomy of query responses was designed after an analysis of 1051 examples of query-query response pairs obtained from the BNC. The sample was obtained from blocks D, F, G, H, J, K of the BNC (so it covers a wide range of dialogue domains, like interviews, radio and TV broadcasts, tutorials, meetings, training sessions or medical consultations). Initially, examples were obtained with the search engine SCoRE (Purver, 2001) (the search string was `?$ | ?$`). Subsequently, cross talk and tag questions were eliminated manually. The sample was classified and annotated by the first author with tags presented in Table 1 (we discuss the reliability of this annotation in section 3).

In what follows we describe and exemplify each class of the resulting taxonomy. To make the description clear we will use $q1$ for the initial question posed and $q2$ for a question given as a response to $q1$. The taxonomy was built with attention paid to the function of $q2$ in the dialogue.

Clarification requests (`CR`) Clarification requests are all question-responses that concern the content or form of $q1$ that was not completely understood. This class contains intended content queries (4a), repetition requests (4b) and relevance clarifications (4c).

(4) a. **A:** What's Hamlet about? **B:** Hamlet? [KPW, 945–946][2]

b. **A:** Why are you in? **B:** What? **A:** Why are you in? [KPT, 469–471]

c. **A:** Is he knocked out? **B:** What do you mean? [KDN, 3170–3171]

In this paper we will not consider this class in detail, mainly because of existing, detailed work on this subject such as (Purver, 2006).

Dependent questions (DP) By a *dependent question* we understand $q2$ where a dependency statement as in (1c) could be assumed to be true. The following examples illustrate this:

(5) a. **A:** Do you want me to <*pause*> push it round? **B:** Is it really disturbing you? [FM1, 679–680]
(cf. *Whether I want you to push it depends on whether it really disturbs you*

b. **A:** Any other questions? **B:** Are you accepting questions on the statement of faith at this point? [F85, 70–71]
(cf. *Whether more questions exist depends on whether you are accepting questions on the statement of faith at this point.*)

'How should I answer this?' questions (FORM) This class consists of question-responses addressing the issue of the *way* the answer to $q1$ should be given. It is the case where the *form* of answer to $q1$ depends on the answer given to $q2$. This relation between $q1$ and $q2$ might be noticed in following examples. Consider (6a). The way B answers A's question in this case will be dictated by A's answer to $q2$—whether or not, A wants to know details point by point.

(6) a. **A:** Okay then, Hannah, what, what happened in your group?
B: Right, do you want me to go through every point? [K75, 220–221]

b. **A:** Where's that one then?
B: Erm, you know Albert Square? [KBC, 506–507]

Requests for underlying motivation (MOTIV) In the case of *requests for underlying motivation* $q2$ addresses the issue of motivation behind asking $q1$. What is important here is that the fact of answering $q1$ depends on the answer to $q2$ (i.e. providing proper reasons for asking $q1$). In this aspect this class differs form the previous ones, where we may assume that an agent wishes to provide answer to $q1$. Most of question-responses of this kind are of the form "Why?" (32 out of 41 gathered examples, cf. (7a)), but also other formulations were observed (8 out of 41, cf. (7b)). Most direct questions of this kind are: *What's it got to do with you?*; *what's it to you?*; *Is that any of your business?*; *what's that gotta do with anything?*.

(7) a. **A:** What's the matter? **B:** Why? [HDM, 470–471]

b. **A:** Out, how much money have you got in the building society? **B:** What's it got to do with you? [KBM, 2086–2087]

I don't want to answer your question (NO ANSW) The role of query responses of this class is to give a clear signal that an agent does not want to provide an answer to $q1$. Instead of answering $q1$ the agent provides $q2$ and attempts to 'turn the table' on the original querier.

(8) a. **A:** Yeah what was your answer? **B:** What was yours? [KP3, 636–637]

b. **A:** Why is it recording me? **B:** Well why not? [KSS, 43–44]

[2] This notation indicates the BNC file (KPW) together with the sentence numbers (945–946).

Table 2: Frequency of question—question-response categories. The parenthesized percentage is the category's percentage of the sample that *excludes* CRs.

Category	Frequency	% of the Total
CR	832	79.16
DP	108	10.28 (49)
MOTIV	41	3.90 (18)
NO ANSW	26	2.47 (12)
FORM	16	1.52 (7)
QA	13	1.24 (6)
IND	9	0.85 (4)
IGNORE	6	0.57 (3)
Total	**1051** (219)	100

Indirect answers/responses (IND/QA) This class consists of query responses, which provide (indirectly) an answer to $q1$. Interestingly, answering $q2$ is not necessary in this case. Consider (9a). B by asking the question *Do you know how old this sweater is?* clearly suggests that the answer to A's question is negative. Moreover, B does not expect to obtain an answer to his/her question. This might also be observed in (9b) ('of course I am Gemini').

(9) a. **A:** Is that an early Christmas present, that sweater? **B:** Do you know how old this sweater is? [HM4, 7–8]

b. **A:** Are you Gemini? **B:** Well if I'm two days away from your, what do you think? [KPA, 3603–3604]

Another means of providing indirect answers can be observed in the corpus data. It is the case that by asking $q2$ an agent already presupposes the answer to $q1$. If we take a look on (10) we note that positive answer to $q1$ is presupposed in B's question (I will help you).

(10) **A:** Will you help with the <*pause*> the paint tonight? **B:** What can I do? [KE4, 3263–3264]

I ignore your question (IGNORE) The last observed class is somewhat harder to grasp. This is the case where $q2$ is related to the situation, but ignores $q1$. This is evident in (11). A and B are playing *Monopoly*. A asks a question, which is ignored by B. It is not that B does not want to answer A's question and that's why he/she asks $q2$. Rather, B ignores $q1$ and asks a question related to the situation (in this case the board game).

(11) **A:** I've got Mayfair <*pause*> Piccadilly, Fleet Street and Regent Street, but I never got a set did I?
B: Mum, how much, how much do you want for Fleet Street? [KCH, 1503–1504]

3 Results and annotation reliability

The results of the performed classification are presented in Table 2. Putting aside CRs, the majoritarian class is indeed DP. What is striking is the relatively large frequency of adversarial responses (the classes MOTIV, NO ANSW, IGNORE). FORM, as we discuss below, is the sole class whose coherence clearly requires reasoning about the querier's intentions. It is relatively infrequent.

In order to check the reliability of the classification process, the decision tree was tested by three other annotators. Annotators obtained the sample of 90 (randomly chosen) question-question pairs[3] and decision tree. The instruction was to annotate question-reply to the first question in each example. Some of the examples were enriched with additional context (after q2). Two annotators reported that the annotation task would be easier if the context would be present for all examples.

The reliability of the annotation was evaluated using κ (Carletta, 1996). The agreement on the coding of the control sample by four annotators was moderate (Fleiss $\kappa = 0.44$, $SE = 0.0206$, $95\%CI =$

[3] The distribution of the classes was in line with the distribution observed by the primary annotator: CR: 39 examples; DP: 18 examples; MOTIV: 10 examples; NO ANSW: 10 examples; FORM: 4 examples; QA: 4 examples; IGNORE: 3 examples; OTHER: 2 examples.

0.3963 to 0.4770)[4]. One of the control sample annotators is an experienced linguist with extensive past work with dialogue transcripts. In this case agreement on the coding was strong (71% agreement with Cohen's $\kappa = 0.62$, $SE = 0.0637$, $95\%CI = 0.4902$ to 0.7398). Two other control sample annotators are logicians, but with little experience in corpus annotation. For them agreement on the coding was somewhat lower, i.e. moderate (66% agreement with Cohen's $\kappa = 0.56$, $SE = 0.0649$, $95\%CI = 0.4266$ to 0.6810; and 54% agreement with Cohen's $\kappa = 0.42$, $SE = 0.0674$, $95\%CI = 0.2829$ to 0.5472). The most unproblematic cases were `CR, MOTIV` and `IGNORE` (the largest groups of examples with at least 3 annotators' agreement). Also `DP, NO ANSW` and `QA` had high agreement between annotators. The main problem was with `FORM`. We assume that this is caused by the unclarity in the question introducing this class in the decision tree ('The way the answer to $q1$ will be given depends on the answer to $q2$', while for DP it was 'Is it the case that the answer to $q1$ depends on the answer to $q2$?'). Feedback from two of three control sample annotators reported this as a confusing case. There were two cases in the control sample on which annotators completely disagreed. These were the following:

(12) a. **A:** You know the one you just took out? **B:** You want it back? [F77, 86–87]

b. **A:** You want a drink dear? **B:** Have your sweets for what? [KD1, 5132–5133]

4 Modeling Query Response Categories in KoS

In this section we show how to explicate the coherence relation that underlies the `DP` query responses within the framework of KoS. It is worth mentioning that this framework allows to model also the other query responses types described in this article, as we will show in an extended version of this paper. KoS is a framework for dialogue whose logical underpinning is Type Theory with Records (TTR) (Cooper, 2005) and which underlies dialogue systems such as GoDiS and CLARIE (Larsson, 2002; Purver, 2006). On the approach developed in KoS, there is actually no single context—instead of a single context, analysis is formulated at a level of information states, one per conversational participant. The type of such information states is given in (13a). We leave the structure of the private part unanalyzed here, as with one exception all our characterizations do not make reference to this; for one approach to $private$, see e.g. (Larsson, 2002). The dialogue gameboard represents information that arises from publicized interactions. Its structure is given in (13b)—the *spkr,addr* fields allow one to track turn ownership, *Facts* represents conversationally shared assumptions, *Pending* and *Moves* represent respectively moves that are in the process of/have been grounded, *QUD* tracks the questions currently under discussion:

(13) a. TotalInformationState (TIS) $=_{def}$

$$\begin{bmatrix} \text{dialoguegameboard : DGB} \\ \text{private : Private} \end{bmatrix}$$

b. DGBType $=_{def}$

$$\begin{bmatrix} \text{spkr: Ind} \\ \text{addr: Ind} \\ \text{utt-time : Time} \\ \text{c-utt : addressing(spkr,addr,utt-time)} \\ \text{Facts : Set(Proposition)} \\ \text{Pending : list(locutionary Proposition)} \\ \text{Moves : list(locutionary Proposition)} \\ \text{QUD : poset(Question)} \end{bmatrix}$$

The basic units of change are mappings between dialogue gameboards that specify how one gameboard configuration can be modified into another on the basis of dialogue moves. We call a mapping between DGB types a *conversational rule*. The types specifying its domain and its range we dub, respectively, the *preconditions* and the *effects*, both of which are supertypes of DGBType.

We start by characterizing the moves that typically involve accepting $q1$ into the DGB. The potential for `DP` responses is explicated on the basis of the two conversational rules in (14a,b): (14a) says that given a question q and ASK(A,B,q) being the LatestMove, one can update QUD with q as QUD–maximal.

[4] All the data established with `http://www.stattools.net`. Access 25.11.2012.

QSPEC is what characterizes the contextual background of reactive queries and assertions. (14b) says that if q is QUD–maximal, then subsequent to this either conversational participant may make a move constrained to be q–specific (14c):

(14) a. Ask QUD–incrementation

$$\begin{bmatrix} \text{pre :} & \begin{bmatrix} \text{q : } & \text{Question} \\ \multicolumn{2}{l}{\text{LatestMove = Ask(spkr,addr,q): IllocProp}} \end{bmatrix} \\ \text{effects :} & \begin{bmatrix} \text{qud} = \left\langle \text{q,pre.qud} \right\rangle \text{: poset(Question)} \end{bmatrix} \end{bmatrix}$$

b. QSPEC

$$\begin{bmatrix} \text{pre : } \begin{bmatrix} \text{qud} = \left\langle \text{q, Q} \right\rangle \text{: poset(Question)} \end{bmatrix} \\ \text{effects : TurnUnderspec } \wedge_{merge} \\ \begin{bmatrix} \text{r : AbSemObj} \\ \text{R: IllocRel} \\ \text{LatestMove = R(spkr,addr,r) : IllocProp} \\ \text{c1 : Qspecific(r,q)} \end{bmatrix} \end{bmatrix}$$

c. q-specific utterance: an utterance whose content is either a proposition p About q or a question q_1 on which q Depends

5 Summary and Future Work

The paper provides the first empirically-based study of query responses to queries. The most interesting finding here is the existence of a number of classes of adversarial responses, that involve the rejection/ignoring of the original query. Indeed, in such cases the original query is rarely responded to in subsequent interaction.

We conducted our study in the BNC since it is a general corpus with a variety of domains and genres. It is of course important to extend this study to more detailed consideration of specific genres and domains. A significant challenge for future work is automatic classification of query responses into a taxonomy like the one provided here. We intend to address this in future work.

References

Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics 22*(2), 249–254.

Carlson, L. (1983). *Dialogue Games*. Synthese Language Library. Dordrecht: D. Reidel.

Cooper, R. (2005). Austinian truth in Martin-Löf type theory. *Research on Language and Computation 3*(4), 333–362.

Ginzburg, J. and R. Fernández (2010). Computational models of dialogue. In A. Clark, C. Fox, and S. Lappin (Eds.), *Handbook of Computational Linguistics and Natural Language*, Oxford. Blackwell.

Ginzburg, J. and I. A. Sag (2000). *Interrogative Investigations: the form, meaning and use of English Interrogatives*. Number 123 in CSLI Lecture Notes. Stanford: California: CSLI Publications.

Larsson, S. (2002). *Issue based Dialogue Management*. Ph. D. thesis, Gothenburg University.

Purver, M. (2001, October). SCoRE: A tool for searching the BNC. Technical Report TR-01-07, Department of Computer Science, King's College London.

Purver, M. (2006). Clarie: Handling clarification requests in a dialogue system. *Research on Language & Computation 4*(2), 259–288.

Rieser, V. and J. Moore (2005). Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Meeting of the Association for Computational Linguistics*, Michigan.

Roberts, C. (2011). Information structure: Afterword. *Semantics and Pragmatics*.

Rodriguez, K. and D. Schlangen (2004). Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In J. Ginzburg and E. Vallduvi (Eds.), *Proceedings of Catalog'04, The 8th Workshop on the Semantics and Pragmatics of Dialogue*, Universitat Pompeu Fabra, Barcelona.

Schegloff, E. (2007). *Sequence Organization in Interaction*. Cambridge: Cambridge University Press.

Wiśniewski, A. (2003). Erotetic search scenarios. *Synthese 134*, 389–427.

Learning Semantic Representations in a Bigram Language Model

Jeff Mitchell
mitchelljeff@hotmail.com

Abstract

This paper investigates the extraction of semantic representations from bigrams. The major obstacle to this objective is that while these word to word dependencies do contain a semantic component, other factors, e.g. syntax, play a much stronger role. An effective solution will therefore require some means of isolating semantic structure from the remainder. Here, the possibility of modelling semantic dependencies within the bigram in terms of the similarity of the two words is explored. A model based on this assumption of semantic coherence is contrasted and combined with a relaxed model lacking this assumption. The induced representations are evaluated in terms of the correlation of predicted similarities to a dataset of noun-verb similarity ratings gathered in an online experiment. The results show that the coherence assumption can be used to induce semantic representations, and that the combined model, which breaks the dependencies down into a semantic and a non-semantic component, achieves the best performance.

1 Introduction

Distributional semantics derives semantic representations from the way that words are distributed across contexts. The assumption behind this approach is that words that occur in similar contexts will tend to have similar meanings. Firth (1957) expressed this in a well known slogan - *you shall know a word by the company it keeps*. In application, these representations have proven successful in automatic thesaurus generation (Grefenstette, 1994), enhancing language models (Coccaro and Jurafsky, 1998) and modelling of reading times (Pynte et al., 2008) and the effects of priming (Landauer and Dumais, 1997).

However, the high level identification of meaning with distributional properties leaves the question of exactly which distributional properties are relevant to semantics a little vague. In practice, researchers evaluate various approaches and select those that produce the best performance. Moreover, other linguistic characteristics, such as syntax, are also analysed in terms of distributional properties. Bigram distributions, for example, are commonly used to induce POS classes (e.g Brown et al., 1992; Clark, 2003), but they have also been investigated as a basis for semantic representations (Bullinaria and Levy, 2007).

Here we examine the question of what statistical properties can be used to distinguish semantic factors from other dependencies in the distribution of words across bigram contexts. We carry this out in terms of class based bigram language models, and explore the possibility that semantic dependencies can be characterised in terms of coherence or similarity across the bigram. We then evaluate the induced representations in terms of their ability to predict human similarity ratings for noun-verb pairs. By evaluating the similarity predictions of our models across POS classes in this way, we assess the ability of the model to focus purely on the semantic content while ignoring other information, such as syntax.

2 Models

The intention is to induce semantic representations within a bigram model based on the assumption that semantic content is coherent across the bigram. Assume that semantic information can be captured in terms of a set, S, of semantic topics, with each word, w, having some independent probability of being used in a topic, $p(w|s)$. Then, if the probabilities of the topics are given by $p(s)$ and each bigram, w_1w_2, belongs to a single topic, then the joint probability, $p(w_1w_2)$, is given by:

$$p(w_1 w_2) = \sum_{s \in S} p(s) p(w_2|s) p(w_1|s) \tag{1}$$

Rewriting this in conditional form, with $p(s|w_1) = \frac{p(s)p(w_1|s)}{p(w_1)}$, gives:

$$p(w_2|w_1) = \sum_{s \in S} p(w_2|s) p(s|w_1) \tag{2}$$

This can also be expressed in a form that explicitly connects to the idea of a probability based on semantic similarity.

$$p(w_2|w_1) = p(w_2) \sum_{s \in S} \frac{p(s|w_2)}{p(s)} p(s) \frac{p(s|w_1)}{p(s)} \tag{3}$$

Equation 3 can be thought of as the unigram probability of w_2 modulated by its similarity to w_1, measured in terms of a weighted dot product between vectors representing the two words. In this case, the vector components are a ratio of probabilities measure, $\frac{p(s|w)}{p(s)}$, which has been widely used in distributional semantics (e.g. Bullinaria and Levy, 2007).

The key feature of this model is that the word probabilities in Equation 1 are independent of position in the bigram. It is this assumption that serves to ensure that the induced topics identify a characterisitic that is stable across the bigram, which, it is hoped, will relate to semantic content.

Relaxing this assumption produces a more general class based model, specifically the aggregate markov model of Saul and Pereira (1997). Using superscripts to indicate the position a word occurs in within the bigram, we write this model as:

$$p(w_2^r|w_1^l) = \sum_{z \in Z} p(w_2^r|z) p(z|w_1^l) \tag{4}$$

In contrast to Equation 1, this model makes no assumption of stability of content across the bigram, and instead allows the word distributions, $p(w|z)$ to be very different in the left and right positions. Thus, this model ought to more suited to handling the word order effects that the similarity based model cannot.

To construct a combined model, the bigram probabilities are expressed in terms of a sum over both S and Z.

$$p(w_2^r|w_1^l) = \sum_{s \in S, z \in Z} p(w_2^r|s, z) p(s, z|w_1^l) \tag{5}$$

These terms can be broken down further based on conditional independence of s and z. The rightmost probability, $p(s, z|w_1^l)$ separates straightforwardly.

$$p(s, z|w_1^l) = p(s|w_1) p(z|w_1^l) \tag{6}$$

On the other hand, s and z cannot in general also be conditionally independent given w_2. However, we can use this as an approximation and then normalise the final probabilities.

$$\hat{p}(w_2^r|s, z) = \frac{p(w_2^r) p(s|w_2) p(z|w_2^r)}{p(s, z)} \tag{7}$$

The final model then combines these components and divides through by a normalising constant $N(w_1)$.

$$p(w_2^r|w_1^l) = \sum_{s \in S, z \in Z} \frac{\hat{p}(w_2^r|s, z) p(s, z|w_1^l)}{N(w_1)} \tag{8}$$

$$N(w_1) = \sum_{w_2} \sum_{s \in S, z \in Z} \hat{p}(w_2^r|s, z) p(s, z|w_1^l) \tag{9}$$

	High	Medium	Low
Group 1	*anticipation - predict* *withdrawal - retire*	*analysis - derive* *invasion - merge*	*opinion - vanish* *disappearance - believe*
Group 2	*disappearance - vanish* *invasion - occupy*	*anticipation - believe* *opinion - predict*	*withdrawal - derive* *implication - retire*
Group 3	*opinion - believe* *implication - derive*	*disappearance - retire* *withdrawal - vanish*	*anticipation - succeed* *invasion - predict*

Table 1: Example items from the noun-verb similarity rating experiment.

2.1 Construction

Models were constructed based on three approaches: similarity based models, as defined by Equation 2, aggregate models, defined by Equation 4, and combined models, defined by Equation 8. The parameters of these bigram models were optimised over a set of sentences extracted from the BNC (BNC Consortium, 2001). 80,775,061 words from the written component of this corpus were used as a training set, with 9,759,769 words forming a development set and the final 9,777,665 words held back as a test set. Preprocessing included conversion to lowercase, addition of $\langle start \rangle$ and $\langle stop \rangle$ at the beginning and ends of sentences, and replacement of words that occurred fewer than 100 times in the training set with an $\langle unk \rangle$ token.

Optimisation of the parameters was based on the EM algorithm (Dempster et al., 1977), with training stopped when the log-likelihood over the development set began to increase. For the pure similarity and aggregate approaches, models were trained with numbers of induced classes ranging from 10 to 2,000. The numbers of classes, $|S|$ and $|Z|$, for the two components of the combined models, each ranged from 10 to 100. The ratio of probabilities measure from Equation 3 was used to construct the components of vectors which then formed the word representations, and similarity of these vectors was measured in terms of the cosine measure.

For comparison, a bigram language model with back-off and Kneser-Ney smoothing (Kneser and Ney, 1995) was also constructed using the SRILM toolkit (Stolcke, 2002).

3 Evaluation

The induced representations were evaluated in terms of their ability to predict semantic similarity ratings for a set of word pairs. We measured the cosine similarity of our word representations and correlated that with the human ratings to produce a measure of agreement. Because the strongest dependencies within the bigrams are likely to be syntactic effects based on the POS classes of the two words, measuring semantic similarity across POS classes is particularly relevant. That is, the semantic representations should contain as much information about the meaning of the words as possible, while containing as little part-of-speech information as possible, which should instead be shifted into the other part of the model. Predicting the similarity between nouns and verbs should therefore be an effective evaluation, as these two word classes contain the core of a sentence's semantic content while having substantially divergent distributional properties in regards of syntax. In this way, we can test whether the POS differences are genuinely being ignored to allow just the semantic similarity to be focussed on.

Thus, an experiment was run to collect similarity ratings for noun-verb pairs. Each participant rated one of three groups of 36 noun-verb pairs, giving a total of 108 items. Each group consisted of 12 high similarity pairs, 12 medium similarity pairs and 12 low similarity pairs.

Table 1 contains a small sample of these items, with rows corresponding to the three experimental groups of participants and columns corresponding to the high, medium and low similarity sets of items seen by each group. The items in the high similarity set (e.g. *anticipation-predict*) are related, via an intermediary word, by a combination of morphology (e.g. *anticipation-anticipate*) and synonymy (e.g. *anticipate-predict*), drawing on Catvar (Habash and Dorr, 2003) and WordNet (Miller, 1995) to identify these relationships. The medium and low sets are then recombinations of nouns and verbs from the high set, with the medium items being the most similar such pairings, as rated by WordNetSimilarity (Pedersen et al., 2004), and low being the least similar.

60 participants were paid \$2 each to rate all 36 items from a single group, with equal numbers

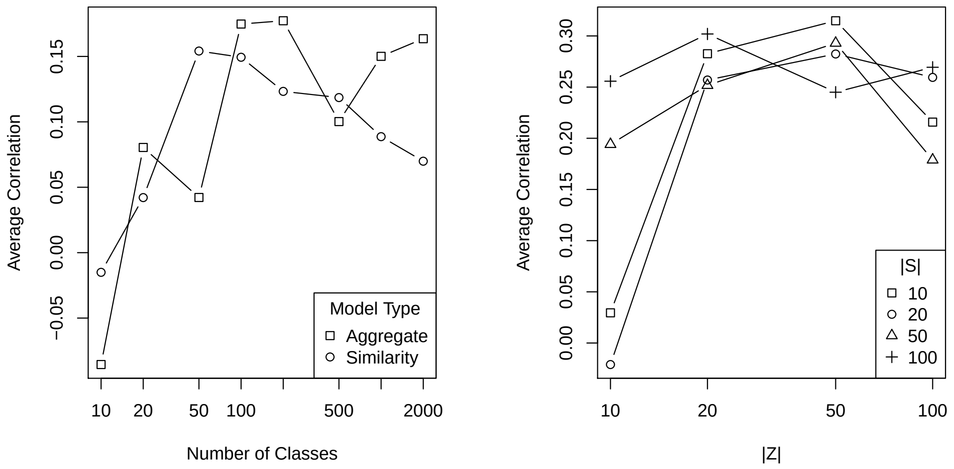
seeing each group. The experiments were conducted online, with participants instructed to rate the similarity in meaning of each pair of words on a scale of 1 to 5. They were initially presented with five practice items before the experimental materials were presented with randomisation of both within and between item orders.

Individual Spearman correlations were calculated for each participant's ratings against the predicted similarities, and the average of these values was used as the evaluation measure for the semantic representations induced by a model. A t-test on the z-transformed participant correlations was used to assess the significance of differences between these averages.

The performance of these models simply as language models was also evaluated, in terms of their perplexity over the test set, T, calculated in terms of the probability assigned to the test set, $p(T)$, and the number of words it contains, $|T|$.

$$perplexity = p(T)^{-\frac{1}{|T|}} \tag{10}$$

4 Results



(a) Average correlations by model size for the Similarity and Aggregate models.

(b) Average correlations by model size for the Combined models.

Figure 1: Correlations of model similarities with human ratings.

Figure 1(a) plots the average correlation between the model similarities and the human ratings for the similarity and aggregate representations. Both models show similar strengths of correlation and a similar pattern in relation to the size of the model, with a peak around the 50 - 200 range. The highest correlation is 0.18, achieved by the aggregate model with 200 classes, while the similarity model achieves a peak of 0.15 at $|S| = 50$. These values are not significantly different, $t(59) = 0.71$, $p = 0.24$. The equivalence in performance of the aggregate and similarity models is not entirely surprising, as both models, despite their differing forms, are directed at the problem of predicting the same bigram dependencies. It may therefore be expected that the weaker semantic factors play only a minor role within the representations generated.

In contrast, the combined models, which allow a separation of the dependencies into distinct components, are able to achieve higher correlations, as plotted in Figure 1(b). Among these models, the highest correlation of 0.31, which is significantly greater than the best aggregate model, $t(59) = 9.35$, $p < 0.001$, is achieved by a model having $|Z| = 50$ and $|S| = 10$. In fact, all the correlations over 0.2 in Figure 1(b) are significantly greater at the $p < 0.001$ level, except $|Z| = 100$, $|S| = 10$ and $|Z| = 20$, $|S| = 20$, which are only significant at the $p < 0.05$ and $p < 0.01$ levels respectively. This leaves only the four lowest performing combined models as not significantly outperforming the best aggregate model. Nonetheless, these values are substantially lower than the inter-subject correlations ($mean = 0.74$, $min = 0.64$), suggesting that the model could be improved further. In particular, extending the span of the model to longer ngrams ought to allow the induction of stronger and more detailed semantic representations. The fact that the best performing model only contains 10 semantic classes underscores the limitations of extracting such representations from bigrams.

In addition to the ability of these models to induce semantic representations, their performance simply as language models was also evaluated. Figure 2 plots perplexity on the test set against number of parameters per word ($|S| + 2|Z|$) for the aggregate and combined models. In general lower perplexities are achieved by larger models for both approaches, as is to be expected. Within this trend, the combined model tends to have a lower perplexity than the aggregate model by about 5%. The single case in which the combined model is above the trend line of the aggregate model occurs for a model with in which a very small aggregate component, $|Z| = 10$, is dominated by a large similarity component, $|S| = 100$.

The performance of these models does not, however, rival that of a standard bigram model with back-off and Kneser and Ney (1995) smoothing, which achieves a perplexity of 185. On the other hand, neither the aggregate nor combined models are explicitly designed to address the issue of small or zero counts.

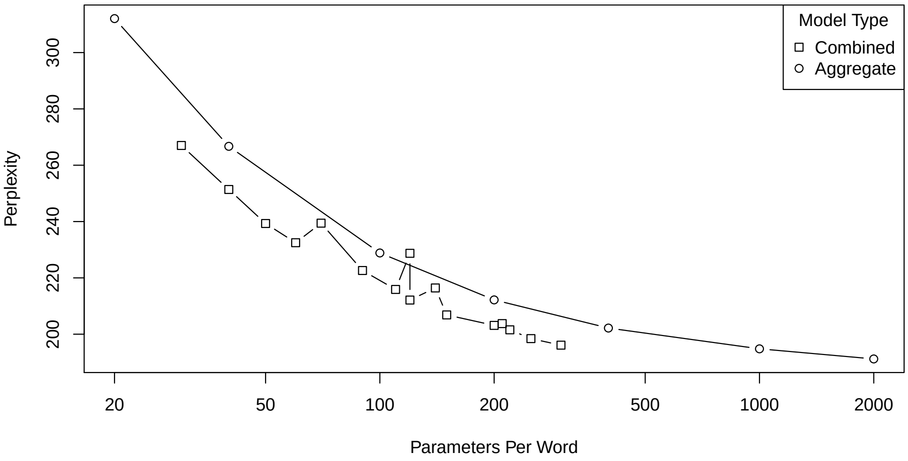


Figure 2: Perplexity by Number of Parameters for the Aggregate and Combined Models

5 Conclusions

Our experiments produced two novel results.

Firstly, we have shown that semantic representations can be induced from the dependencies within bigram word sequences, using an approach that derives word probabilities from similarity. This is similar in form to prior models (e.g Coccaro and Jurafsky, 1998; Bellegarda, 1997), but whereas they imported distributional representations from outside the model to enhance their performance, we use this model form to induce semantic representations within the model.

Secondly, we have shown that this approach is most effective when the model breaks these dependencies down into both a semantic and a non-semantic component. Typically, semantic classes have been induced in isolation (Landauer and Dumais, 1997; Bullinaria and Levy, 2007) or applied to long-range structure while short-range structure is handled by a separate component (Boyd-Graber and Blei, 2008; Griffiths et al., 2004; Wallach, 2006). Here, we have shown that even simple bigram dependencies can be conceived as breaking down into semantic and non-semantic components, as opposed to using those components to model two different types of dependency.

We also introduced a novel evaluation dataset for semantic representations, containing noun-verb similarity ratings. Correlation of these human ratings with the model similarities allows a quantification of the extent to which a model ignores POS information to focus on semantic content.

In future work, we hope to extend the span of our model and to characterise syntax, semantics and their interaction in a more sophisticated manner. Particularly interesting is the question of the extent to which the form of our model is specific to languages, such as English, in which syntax is identified with word order and how this might be adapted to free word order languages.

References

Bellegarda, J. R. (1997). A latent semantic analysis framework for large-span language modeling. In G. Kokkinakis, N. Fakotakis, and E. Dermatas (Eds.), *EUROSPEECH*. ISCA.

BNC Consortium (2001). The British National Corpus, Version 2. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Boyd-Graber, J. L. and D. M. Blei (2008). Syntactic topic models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), *NIPS*, pp. 185–192. Curran Associates, Inc.

Brown, P. F., V. J. D. Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer (1992). Class-based n-gram models of natural language. *Computational Linguistics 18*(4), 467–479.

Bullinaria, J. and J. Levy (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods 39*(3), 510–526.

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *EACL*, pp. 59–66. The Association for Computer Linguistics.

Coccaro, N. and D. Jurafsky (1998). Towards better integration of semantic predictors in statistical language modeling. In *ICSLP*. ISCA.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1–32. Oxford: Philological Society.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Griffiths, T. L., M. Steyvers, D. M. Blei, and J. B. Tenenbaum (2004). Integrating topics and syntax. In *NIPS*.

Habash, N. and B. J. Dorr (2003). A categorial variation database for english. In *HLT-NAACL*.

Kneser, R. and H. Ney (1995). Improved backing-off for M-gram language modeling. In *International Conference on Acoustics, Speech, and Signal Processing, 1995.*, Volume 1, pp. 181–184 vol.1.

Landauer, T. and S. Dumais (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review 104*(2), 211.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM 38*(11), 39–41.

Pedersen, T., S. Patwardhan, and J. Michelizzi (2004). Wordnet: : Similarity - measuring the relatedness of concepts. In D. L. McGuinness and G. Ferguson (Eds.), *AAAI*, pp. 1024–1025. AAAI Press / The MIT Press.

Pynte, J., B. New, and A. Kennedy (2008). On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research 48*, 2172–2183.

Saul, L. K. and F. Pereira (1997). Aggregate and mixed-order markov models for statistical language processing. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, New York, NY, pp. 81–89. ACM Press.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In J. H. L. Hansen and B. L. Pellom (Eds.), *INTERSPEECH*. ISCA.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In W. W. Cohen and A. Moore (Eds.), *ICML*, Volume 148 of *ACM International Conference Proceeding Series*, pp. 977–984. ACM.

Predicate-specific Annotations for Implicit Role Binding: Corpus Annotation, Data Analysis and Evaluation Experiments

Tatjana Moor Michael Roth Anette Frank
Department of Computational Linguistics, Heidelberg University
{moor,mroth,frank}@cl.uni-heidelberg.de

Abstract

Current research on linking implicit roles in discourse is severely hampered by the lack of sufficient training resources, especially in the verbal domain: learning algorithms require higher-volume annotations for specific predicates in order to derive valid generalizations, and a larger volume of annotations is crucial for insightful evaluation and comparison of alternative models for role linking.

We present a corpus of predicate-specific annotations for verbs in the FrameNet paradigm that are aligned with PropBank and VerbNet. A qualitative data analysis leads to observations regarding implicit role realization that can guide further annotation efforts. Experiments using role linking annotations for five predicates demonstrate high performance for these target predicates. Using our additional data in the SemEval task, we obtain overall performance gains of 2-4 points F_1-score.

1 Introduction

Automatic annotation of semantic predicate-argument structure (PAS) is an important subtask to be solved for high-quality information access and natural language understanding. Semantic role labeling (SRL) has made tremendous progress in addressing this task, using supervised and recently also semi- and unsupervised methods (Palmer et al., 2010).

Traditional SRL is restricted to the local syntactic domain. In discourse interpretation, however, we typically find locally unrealized argument roles that are contextually bound to antecedents beyond their local structure. Thus, by using strictly local methods, we are far from capturing the full potential offered by semantic argument structure (Fillmore and Baker, 2001; Burchardt et al., 2005).

The task of resolving the reference of implicit arguments has been addressed in previous work: Gerber and Chai (2012) address the task in the nominal domain by learning a model from manually annotated data following the NomBank paradigm (Meyers et al., 2004). In contrast, Ruppenhofer et al. (2010) follow the FrameNet paradigm, which is not restricted to nominal predicates. However, their data set suffers from considerable sparsity with respect to annotation instances per predicate (cf. Section 2).

Our contribution addresses the problem of sparse training resources for implicit role binding by providing a higher volume of predicate-specific annotations for non-local role binding, using OntoNotes (Weischedel et al., 2011) as underlying corpus. A qualitative analysis of the produced annotations leads to a number of hypotheses on implicit role realization. Using the extended set of annotations, we perform experiments to measure their impact, using a state-of-the-art system for implicit role binding.

2 Motivation and Related Work

The main motivation for this work relates to the SemEval 2010 Task 10 on implicit role linking[1] and the problem of data sparsity that became evident by the poor performance of the participating systems, at 1% F_1-score (Tonelli and Delmonte, 2010; Chen et al., 2010).[2] Later systems could only marginally improve

[1] http://www.coli.uni-saarland.de/projects/semeval2010_FG

[2] The data set provides 245/259 instances of resolvable implicit roles for training/testing. All cases of implicit roles (580/710) are distributed over 317/452 frame types and a small overall number of frame instances (1,370/1,703 training/testing).

on these results, with performance up to 19% F_1-score due to improved recognition of resolvable implicit roles, heuristic data acquisition, and variations in model properties.[3] Gerber and Chai (2010, 2012), working on a related task following the NomBank/PropBank paradigm, achieved higher performance of 50% F_1-score, using as training data a substantial amount of annotations for 10 noun predicates.

3 Corpus and Annotation

3.1 Corpus

While there is a rich source of annotated sentences in the FrameNet paradigm, contextualized FrameNet annotations are restricted in coverage. As we target high-frequency annotations for specific verbs, and in order to make annotations available for a corpus that is widely used, we chose OntoNotes (V.4.0) (Weischedel et al., 2011) as underlying corpus. OntoNotes contains semantic role annotations following the PropBank annotation style (Palmer et al., 2005). We map these annotations to FrameNet using the mapping specified in the SemEval 2010 Task 10 data, which is based on SemLink (Loper et al., 2007).

3.2 Selection of Annotation Targets

Our goal was to produce a high volume of annotations for specific verb predicates, ideally reaching a margin of 100-200 instances involving locally unfilled argument roles (cf. Gerber and Chai (2010)). In order to make the task feasible for the annotators, we selected predicates and frame readings that are relatively easy to discriminate, so that the annotators can concentrate on the role linking task.

We applied a number of further selection criteria to make the resulting annotations as useful as possible: (i) We excluded light verbs, as they are not well covered in FrameNet, and typically involve difficult sense distinctions. (ii) We only chose predicates (and senses) that are covered in VerbNet, PropBank and FrameNet, according to the Unified Verb Index.[4] This ensures that the corpus can also be used for experimentation using the VerbNet or PropBank paradigm. Finally, (iii) for the selected candidate predicates and readings, we investigated the FrameNet annotation data base to determine whether the annotated frames involve a critical number of non-instantiated roles that can be resolved from discourse context.[5] In case we found little or no such cases for the candidate reading, the predicate was not chosen.

The list of predicates that resulted from this selection process is given in Table 1. They exhibit varying numbers of core roles (2-7), frame ambiguity (1-7), and different syntactic properties.

3.3 Annotation Process and Categories

Data preparation. We extracted annotation instances for the selected target predicates from the OntoNotes corpus. The resulting corpus consists of overall 1.992 instances. Each annotation instance was embedded within its full document context. The average document length is 612 words.

Annotation Categories. Our annotation maily follows the SemEval task guidelines for role linking (Ruppenhofer et al., 2010, 2012), with the exception that we differentiate between non-instantiated (NI) roles that are *resolvable* vs. *non-resolvable* within discourse instead of classifying them as ‘definite (DNI)’ vs. ‘indefinite (INI)’. This distinction makes the task of linking NIs much clearer, as definite null-instantiations may or may not be resolvable within the discourse context.[6]

Two examples of NI occurrences are given below: (1) illustrates a (resolvable) DNI: the implicit role’s referent is anaphorically bound within the prior discourse. In (2) the non-instantiated role can only be interpreted existentially within the given discourse (non-resolvable, INI).

[3] See Tonelli and Delmonte (2011); Ruppenhofer et al. (2011); Silberer and Frank (2012); Laparra and Rigau (2012).

[4] `http://verbs.colorado.edu/verb-index`

[5] Even though FrameNet annotations are out of context, non-realized core roles are marked for definite vs. indefinite interpretation.

[6] In fact, only 80.9%/74.2% of all DNIs in the SemEval training/test corpus are linked within discourse.

(1) (s3) Nearly 200 American agents went to $[\text{Yemen}]_{Source}$ right after the attack on the "Cole".
(s9) They_{Theme} are *leaving* frustrated. (*Source*: resolvable, DNI)

(2) I_{Donor} tried to *give* as good as I got. (*Recepient*, non-resolvable, INI; *Theme*, non-resolvable, INI)

The annotation consists of three sub-tasks that are applied to null-instantiated core roles (NIs) only: (a) classifying each NI as a *resolvable or non-resolvable null instantiation*; (b) distinguishing between *Lexical* and *Constructional Licensing* of each NI;[7] and (c) *linking resolvable DNIs* to the closest antecedent mention within the preceding context.[8]

Before proceeding to these decisions, the annotator determines whether the mapped frame corresponds to the actual predicate meaning in the given context. If not, it is marked as 'NgFNS' (no genuine FN sense). We also flag roles whose filler does not correspond to the role definition given in FrameNet (e.g., roles categorized as 'Physical object' that are filled by an abstract concept). For each predicate, we record the chosen frame as well as the mapping to the corresponding readings in PropBank and VerbNet.

Calibration of Annotation Quality. The annotation was performed by two annotators, both students of Computational Linguistics. Both of them studied the SemEval guidelines and used the first 50 sentences of the SemEval corpus as a trial corpus, in order to validate their understanding of the guidelines.

We performed two calibration experiments, in which we measured Kappa (Cohen, 1960) for the assignment of role interpretation type (resolvable vs. non-resolvable role), and percentage of agreement on the chosen antecedent for resolvable roles.

(I) **Agreement with SemEval:** After initial training, we measured IAA between our main annotator and the SemEval gold annotations for sentences 51-100 of the SemEval data set. For *interpretation type* (resolvable/non-resolvable classification) we achieved a Kappa of 0.77. For NI-linking, we measured 71.43% agreement (15 out of 21 resolvable roles were correctly linked).

(II) **Agreement between Annotators:** We determined agreement between both annotators on all annotation instances pertaining to the predicate *give*. For *interpretation type*, the annotators achieved a Kappa value of 0.94. For *linking*, the annotators agreed on the marked antecedent in 85.7% of all cases (48 out of 56 cases).

After this calibration phase, the annotation was done independently by the two annotators.

4 Data Analysis

Table 1 gives an overview of the annotations we obtained. Overall we annotated 630 NI occurrences for genuine frame meanings, distributed over 438 verb occurrences (i.e., 1.44 NIs/verb).[9] We observe great variation in the number of NI occurrences for the different predicates (e.g., *leave* vs. *pay*). We find a predominance of non-resolvable over resolvable role classifications (61.6% vs. 38.4%). 78% of the resolvable NIs are realized within a window of 3 sentences,[10] as opposed to 69.6% in the SemEval and 90% in Gerber&Chai's data. This can be explained by variation in text genre and target categories.

Considering the distribution of NI-realizations and the properties of the corresponding predicates, we note some tendencies that seem worth investigating on a larger scale, as potential factors determining null instantiation of roles. Predicates with low frame ambiguity rate (*pay, bring, give*) tend to have a higher NI-realization rate than frames with a higher ambiguity rate (*leave, put*). A higher number of core roles of the target frame tends to go along with a higher NI-realization potential (*bring, pay*).

[7] As the SemEval guidelines for lexical and constructional licensing are not very explicit, and given these annotations are not required for evaluating system annotations, we do not report details about this part of the annotation.

[8] If the antecedent is not found in the preceding context, we also inspect the following discourse.

[9] 204 out of all 834 NIs (24.5%) do not pertain to genuine frame readings. These were held out from further data statistics.

[10] 9.9% of the fillers were found in the following discourse.

verb	frame	verb	core	frame	NI occurrences								
		occ.	roles	ambig.	all	other	frame	frame	NIs per	resolvable		non-resolv.	
						reading	reading	occ.	frame	abs.	%	abs.	%
give	GIVING	524	3	1	218	63	155	144	1.08	62	40.0	93	60.0
put	PLACING	427	4	3	39	17	22	22	1.00	10	45.5	12	54.5
leave	DEPARTING	354	2	7	70	30	40	39	1.03	25	62.5	15	37.5
bring	BRINGING	351	7	2	103	38	65	45	1.44	28	43.1	37	56.9
pay	COMMERCE_PAY	336	5	1	404	56	348	188	1.85	117	33.6	231	66.4
all		1992	–	–	834	204	630	438	–	242	–	388	–

Table 1: Annotated predicates and data analysis: Implicit role interpretation and linking.

give: GIVING		all	**Donor**	**Recepient**	**Theme**		
Interpretation	resolvable	40.0	**25.2**	13.5	1.3		
	non-resolvable	60.0	19.4	**37.4**	3.2		
put: PLACING		all	**Agent**	**Cause**	**Theme**	**Goal**	
Interpretation	resolvable	45.5	**40.9**	4.6	0.0	0.0	
	non-resolvable	54.4	**45.5**	9.1	0.0	0.0	
leave: DEPARTING		all	**Theme**	**Source**			
Interpretation	resolvable	62.5	0.0	**62.5**			
	non-resolvable	37.5	7.5	30.0			
bring: BRINGING		all	**Agent**	**Goal**	**Source**	**Carrier**	
Interpretation	resolvable	43.1	9.3	16.9	16.9	0.0	
	non-resolvable	56.9	21.5	1.5	**23.1**	10.8	
pay: COMMERCE_PAY		all	**Buyer**	**Seller**	**Goods**	**Money**	**Rate**
Interpretation	resolvable	33.6	6.6	**14.6**	9.2	2.9	0.3
	non-resolvable	66.4	2.5	**25.3**	17.2	10.9	10.3

Table 2: Distribution of resolvable vs. non-resolvable NI roles over predicate roles (in percent).

Further data statistics for particular predicates and which roles they omit under the different NI-interpretations are given in Table 2. Typically, we find NI-realization concentrated on one or two roles for a given predicate. Yet, these are observations on a small number of predicates that need substantiation by further data annotation, with systematic exploration of other determining properties, such as role meaning or perspectivation (*pay/sell*; *leave/arrive*) and the influence of constructional licensing.

5 Evaluation Experiments

We evaluate the impact of predicate-specific annotations for classification using two scenarios: (**CV**) we examine the linking performance of models trained and tested on the same predicate by adopting the 10-fold Cross-Validation scenario used by Gerber and Chai (2012) (G&C).[11] (**SemEval**) Secondly, we examine the direct effect of using our annotations as additional training data for linking NIs in the SemEval 2010 task on implicit role binding. We use the state-of-the-art system and best performing feature set described in Silberer and Frank (2012) (S&F) to make direct comparisons to previous results.

CV. As positive training and test samples for this scenario, we use all annotated (and resolvable) NIs and randomly split them into 10 folds. Negative training samples (i.e., incorrect NI fillers) are automat-

[11] Note that this is not a direct comparison as both the annotation paradigm and the data sets are different.

Cross-Validation				SemEval 2010 test set					
verb	precision	recall	F_1	**training data**	FS	NgFNS	precision	recall	F_1
give	48.8	33.3	39.6	**S&F'12**					
put	33.3	20.0	25.0	no additional data	+	n.a.	25.6	25.1	25.3
leave	48.3	**56.0**	**51.9**	+ best heuristic data	+	n.a.	30.8	25.1	27.7
bring	**72.7**	27.6	40.0	**this paper**					
pay	35.4	20.0	25.6	+ our annotations	–	–	21.7	21.2	21.5
—	—	—	—	+ our annotations	+	–	33.3	22.0	26.5
average	47.7	31.4	36.4	+ our annotations	+	+	**34.3**	**26.3**	**29.8**

Table 3: Results for both evaluations (all figures are percentages). FS indicates whether feature selection was applied. NgFNS indicates the use of frame annotations that do not match the contextual meaning.

ically added by extracting constituents that overtly fill a role according to the semantic annotations in the OntoNotes gold standard. We only consider phrases of type NPB, S, VP, SBAR and SG within the current and the two preceding sentences as potential fillers.[12]

SemEval. This setting is identical with the linking evaluation in S&F. Like them, we (optionally) apply an additional step of feature selection (±FS) on the SemEval training data to select a feature subset that generalizes best across data sets, i.e., the fully annotated novel from the shared task and our predicate-specific annotations based on OntoNotes. We further compare models trained w/ and w/o non-genuine frame annotations (±NgFNS). As in the CV setting, we assume that all resolvable NIs are known and only the correct fillers are unknown. Thus, our results are not comparable to those of participants of the *full* SemEval task, who solved two further sub-tasks. Instead we compare to the NI linking results in S&F, with models trained on the SemEval data and using additional heuristically labelled data.

Table 3 summarizes our results for both settings. They are not strictly comparable due to varying properties, i.a., the number of available annotations. The **CV** results show that few annotations can be sufficient to achieve a high linking precision and f-score (up to 72.7 P, 51.9 F_1). However, this is highly dependent on the target predicate (cf. *bring* vs. *pay*). Overall, the results exhibit a similar variance and lie within the same range as those reported by G&C. Even though the numbers are not directly comparable, they generally indicate a similar difficulty of linking implicit arguments across lexical predicate types.

In the **SemEval** setting, we obtain improved precision and recall over S&F's results (± additional heuristic data, cf. Silberer&Frank, 2012)) when linking NIs using our additional training data and feature selection. Using our full additional data set (+NgFNS) we obtain higher performance compared to S&F's best setting with heuristically labelled data, yielding highest scores of 34.3% precision and 26.3% recall. The resulting F_1-score of 29.76% lies 2.1 points above the best model of S&F, whose full system also achieved state-of-the-art performance on the *full* SemEval task.

6 Conclusions

We presented an annotation effort for implicit role linking targeting five verb predicates. The FrameNet annotations are mapped to PropBank and VerbNet and will be available for the community. Annotations follow the SemEval guidelines and were quality-controlled. We annotated 630 NI realizations for the intended predicate senses. Our experiments show that even a moderate amount of annotations per predicate yield substantial performance gains of 2.1-4.5 points F_1-score. Our data set complements the SemEval corpus in terms of text genre and Gerber&Chai's data set in terms of category and explicit annotation for interpretation type. Due to higher-volume predicate-specific annotations, it enables more insightful evaluation and comparison between different models, including comparison across frameworks. In future work, we plan to extend the annotation to further predicates using, i.a., active learning techniques.

[12] This corresponds to the *SentWin* setting in Silberer and Frank (2012) and is motivated by the fact that most NI fillers both in the SemEval training data and in our annotations are located within a span of the current and two preceding sentences.

References

Burchardt, A., A. Frank, and M. Pinkal (2005). Building Text Meaning Representations from Contextually Related Frames – A Case Study. In *Proceedings of the 6th International Workshop on Computational Semantics*, IWCS-6, Tilburg, The Netherlands, pp. 66–77.

Chen, D., N. Schneider, D. Das, and N. A. Smith (2010, July). SEMAFOR: Frame Argument Resolution with Log-Linear Models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, pp. 264–267.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37–46.

Fillmore, C. J. and C. F. Baker (2001, June). Frame Semantics for Text Understanding. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh.

Gerber, M. and J. Chai (2010). Beyond NomBank: A Study of Implicit Arguments for Nominal Predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 1583–1592.

Gerber, M. and J. Y. Chai (2012). Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics 38*(4), 755–798.

Laparra, E. and G. Rigau (2012). Exploiting Explicit Annotations and Semantic Types for Implicit Argument Resolution. In *6th IEEE International Conference on Semantic Computing (ICSC'12)*, Palermo, Italy.

Loper, E., S. Yi, and M. Palmer (2007). Combining Lexical Resources: Mapping between PropBank and VerbNet. In *Proceedings of the 7th International Workshop on Computational Semantics*, Tilburg, the Netherlands.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman (2004). Annotating Noun Argument Structure for NomBank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC-2004, Lisbon, Portugal, pp. 803–806.

Palmer, M., D. Gildea, and P. Kingsbury (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics 31*(1), 71–106.

Palmer, M., D. Gildea, and N. Xue (2010). *Semantic Role Labeling*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Ruppenhofer, J., P. Gorinski, and C. Sporleder (2011). In Search of Missing Arguments: A Linguistic Approach. In *Proceedings of RANLP*, Hissar, Bulgaria, pp. 331–338.

Ruppenhofer, J., R. Lee-Goldman, C. Sporleder, and R. Morante (2012). Beyond sentence-level semantic role labeling: linking argument structures in discourse. *Language Resources and Evaluation*. DOI: 10.1007/s10579-012-9201-4.

Ruppenhofer, J., C. Sporleder, R. Morante, C. Baker, and M. Palmer (2010). SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluations*, Uppsala, Sweden, pp. 45–50.

Silberer, C. and A. Frank (2012). Casting Implicit Role Linking as an Anaphora Resolution Task. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, Montréal, Canada, pp. 1–10.

Tonelli, S. and R. Delmonte (2010). VENSES++: Adapting a Deep Semantic Processing System to the Identification of Null Instantiations. In *Proceedings of the 5th International Workshop on Semantic Evaluations*, Uppsala, Sweden, pp. 296–299.

Tonelli, S. and R. Delmonte (2011). Desperately Seeking Implicit Arguments in Text. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, Portland, pp. 54–62.

Weischedel, R., E. Hovy, M. Palmer, M. Marcus, R. Blevin, S. Pradhan, L. Ramshaw, and N. Xue (2011). OntoNotes: A Large Training Corpus for Enhanced Processing. In J. Olive, C. Christianson, and J. McCary (Eds.), *Handbook of Natural Language Processing and Machine Translation*. Springer.

A Pilot Experiment in Knowledge Authoring as Dialogue[*]

Artemis Parvizi[†] Caroline Jay[‡] Christopher Mellish[†] Jeff Z. Pan[†] Yuan Ren[†]
Robert Stevens[‡] Kees van Deemter[†]

Abstract

This project aims to build an ontology authoring interface in which the user is engaged in a dialogue with the system in controlled natural language. To investigate what such a dialogue might be like, a layered annotation scheme is being developed for interactions between ontology authors and the Protégé ontology authoring environment. A pilot experiment has been conducted with ontology authors, which reveals the complexity of mapping between user-interface actions and acts that appear in natural language dialogues; it also suggests the addition of some unanticipated types of dialogue acts and points the way to some possible enhancements of the authoring interface.

1 Introduction

Ontology authoring – the process of creating and modifying an ontology in order to capture the knowledge about a domain – is hard. Studies such as (Rector et al., 2004) and (Dzbor et al., 2006), for example, have shown that ontology authors frequently misunderstand axioms in the ontology. These studies also suggest that current ontology authoring tools fail to support some of the actions that authors would like to perform, and that users would like to be warned against potential mis-uses of axioms and unforeseen consequences of those axioms.

This paper reports on work in progress, in which we study the interaction of human users with an ontology authoring interface, with the ultimate aim of developing a tool that will permit a much richer knowledge authoring experience, thereby addressing the above-mentioned problems associated with existing knowledge authoring. We envisage developing a plugin to the knowledge authoring interface Protégé[1] that will allow authors to interact with the ontology via a controlled natural language (CNL) dialogue that offers users some of the freedom of natural language without causing insurmountable natural language understanding problems to the system. Specifically, we report on a pilot experiment in which we observed human editors who were invited to use Protégé while talking to an experimenter, and on a layered annotation scheme that we are developing for annotating the resulting interactions. Once a stable annotation scheme has been reached, we shall use the scheme to obtain annotations involving a larger number of users and a larger number of knowledge authoring tasks, the results of which will inform the design of the new knowledge authoring interface. In this paper, we focus on the following questions:

- If knowledge authoring is viewed as a dialogue, what would be the main moves that one would expect to see in these dialogues?
- How are these dialogue moves grounded in lower-level actions like "Finding subsumers for a given concept", or "Looking at a concept"?
- How does the annotation of a knowledge authoring dialogue compare to the annotation of real spoken dialogue?

[*] This work was supported by EPSRC grants EP/J014354/1 and EP/J014176/1.

[†] University of Aberdeen, UK

[‡] University of Manchester, UK

[1] http://protege.stanford.edu

Ontology authoring aided by CNL has been addressed from various points. ACE (Fuchs et al., 1999) and PENG (Schwitter, 2002) allow users to express specifications that can be translated into an unambiguous logical language, understandable to machines. CLCE (Sowa, 2004) resembles ACE but, being closer to English, its output is more readable. SWAT (Power, 2009) uses natural language generation to enable users to produce description logic statements. Similar to CLCE, (Bernardi et al., 2012)'s work is another example of a novel controlled language capable of assisting ontology authoring.

Whereas this previous work has addressed the problem of producing isolated utterances relevant to knowledge authoring using CNL, this project attempts to further develop knowledge authoring by adopting new interactive methods inspired by human dialogue. For example, we hypothesise that by allowing an author to pose *what-if* questions prior to authoring an axiom in an ontology, the authoring process will run in a more informed manner. The following dialogue is an example of the type of communication that this project is attempting to make possible during knowledge authoring:

U1 I want a Moral Right to be able to have a date.
S1 Moral Rights currently have no date property.
U2 What if I import timeframe.owl?
S2 That is consistent. There is now a date property.
U3 What if a Moral Right must have exactly one date?
S3 That implies that Attribution Rights and Integrity Rights must have exactly one date.

2 A new dialogue annotation scheme for Knowledge Authoring

We decided to use a layered model of human-computer interaction to analyse the interactions between knowledge editors and Protégé. This general idea has a long pedigree. Moran (1981) proposed a command language grammar, Foley et al. (1982) suggested a 4-layer design model, Buxton (1983) expanded Foley et al.'s model into a 5-layer design model, and Nielsen (1986) presented a virtual communication protocol. Generally, the aim is to have a better understanding of various aspects of human machine interaction, recognising that a high-level action may be "implemented" in various ways. In this section, we sketch our intial three-level scheme for analysing the interaction between users and Protégé, which was then modified in light of our experiment.

KLM level. The lowest level of analysis is akin to the Keystroke Level Model (KLM) (Card et al., 1986). This level records the physical interactions of the user with the system. KLM level includes acts such as clicks, mouse hovers, and eye movements.

Authoring level. At a higher level of analysis, we wanted to record what it is that users attempt to do and the high-level actions that they perform to accommodate these goals. Some existing analyses (Goncalves et al., 2011; Abgaz et al., 2012) have attempted to understand and categorise changes to the knowledge base during its development. The intention of the present work is not only to categorise such change, but also to understand users' thinking and, ultimately, to work towards a greater knowledge authoring functionality. We decided to group authoring-level acts into general acts and interface acts. *General acts* are grouped into (1) various types of observations (for example, looking whether a class of a given name exists in the ontology; looking to see what instances of a class exist), (2) addition, updating, and deletion of axioms, (3) noticing and resolving inconsistencies. *Interface acts* include (1) common interface features such as undoing or redoing a previous action, and (2) seeking explanations of inconsistencies.

Dialogue level. Annotation at this level involves a creative step where actions are analysed *as if* they were instances of acts in a natural language dialogue. These dialogue acts will provide basis for the design of an interface for knowledge authoring that is similar to natural dialogue. Annotation of dialogue acts is standard in the analysis of Natural Language dialogues (Bunt, 1994; Ginzburg, 1994), but this is not yet used in ontology authoring. Figures 1 and 2 show our initial intuition of dialogue acts, which are categorised into user and system's dialogue acts.

- *Informative Questions* The aim of this type of question is to acquire some information, for example, "Where does class X appear in the classification hierarchy?". The system is intended to search

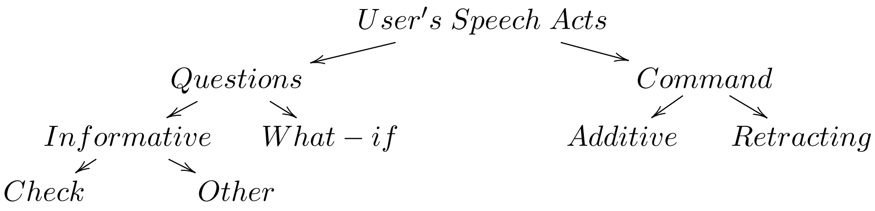


Figure 1: User's Dialogue Acts

through the inferred hierarchy for a response.

- *What-if Questions* Based on the current context, this type of question informs the user about a possible future state of the system. The system needs to perform an action and report its consequences. However, the decision to apply the change is made by the user later.
- *Check Questions* Similar in nature to informative questions, check questions provide answers to user's inquiries, but where the user has a strong expectation about the answer.
- *Additive Command* Informing the system of a certain statement that has to be added or made true in the ontology. The consistency of the ontology needs to be checked and reported to the user. The decision how and when to address the inconsistency has to be made by the user later.
- *Retracting Command* In certain circumstances when the ontology has become inconsistent, the user can either ask an informative question, or a check question and based on the answer decide to retract a change; or, the user can simply retract one or more of the uttered statements.

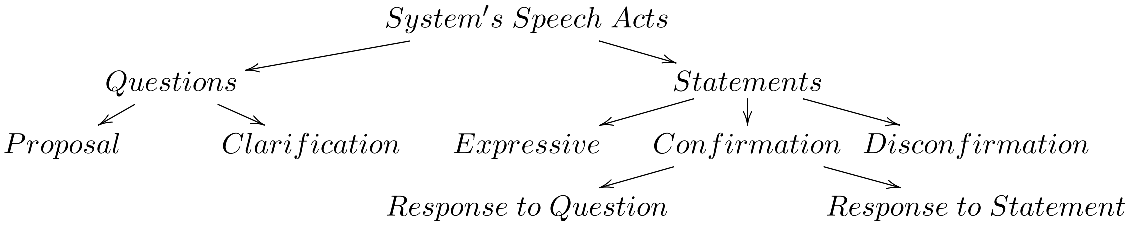


Figure 2: System's Dialogue Acts

- *Expressive Statements* A response to an information seeking or a *what-if* question. Based on the type of question to be addressed, the system may or may not be involved in an authoring process.
- *Confirmation Statements in Response to Questions* Informing the user that an authoring act had the desired effect.
- *Disconfirmation Statements in Response to Question* Informing the user of the consequences to an authoring act, which may not be desired.
- *Confirmation Statements in Response to Statements* Confirmation of a successful authoring act.
- *Proposal Questions* A suggestion by the system to undertake an action.
- *Clarification Questions* A question posed to the user to clarify a command.

3 A Pilot Experiment

The 5 participants in the pilot study were computer scientists who were experienced users of OWL and Protégé. They were asked to set up the Protégé interface as they normally operated it. They were asked to explore the People and Pet ontology acquired from the Tones repository [2], and manipulate it as they saw fit. They were asked to follow the *think aloud* protocol, and describe their moves. In addition, their interactions and eye movements were video-recorded. Where the experimenters were unclear about a description of what was happening, they asked the subject for clarification. Occasionally, when the

[2] http://owl.cs.manchester.ac.uk/repository/

subject felt unsure how to proceed, the experimenter suggested an action (for example, Why don't you look at the axioms now appearing in red?")

Table 1: A pilot experiment and various levels of annotation

	Speech/Action	**Authoring level**	**Dialogue Level**
7:35	"he's a car driver"	Adopt goal 1: (Mick is a car driver)	Expression of intention
⋮	⋮	⋮	⋮
8:57	"make the . . . thing that he's driving . . . "	Adopt goal 2: (Q123 is a car)	Expression of intention
9:01	"I will just remove whatever was there and . . . "	Adopt goal 3: (Q123 is only a car)	Expression of intention
9:03	"oh I know!"	Abandon goal 3	Retraction of intention
9:03	clicks 'Types+' in 'Description Q123'	Add axiom: (Q123 is a car)	Additive command
9:04	"add car as well"		
9:04	types 'car' in 'Q123' dialogue		
9:05	"and see whether it's invalid"	Check: (Mick is a car driver)	Check Question
9:08	runs reasoner		
9:08	"if it is really the same thing then it should be disjoint"	Notice: (vehicle type not disjoint)	–
		Note goal: (vehicle types disjoint)	Expression of desire
9:11	"mm-hm"	Confirm check	Confirmation statement
9:16	clicks 'Mick' in 'individuals' window	Goal achieved: (Mick is a car driver)	Retraction of intention

Knowledge authoring sessions were annotated with reference to the scheme outlined above but, crucially, annotators wrote annotations in a free style: the interface did not limit them to a fixed set of labels implied by our initial annotation scheme. A small annotated segment of one session is presented in Table 1. In the speech/action column, speech refers to the user's comments and is displayed within quotations, and actions refers to the physical interactions between the user and the Protégé interface.

Our initial intuitions about dialogue acts were shown to be far from perfect. Perhaps unsurprisingly, it became clear that the mapping between KLM, Protégé and dialogue levels can be very complex. An annotation in one layer can correspond to more than one in another layer. For example, a subject may observe the definition or the location in the hierarchy of a concept for various *conceptual reasons*, for instance to check consistency (a *Check Question* in our scheme) or to prepare for adding a new concept (an *Additive Command*). Difficulties of a similar kind are well known from the annotation of natural language dialogue, for instance whether a given spoken utterance (for example, "You are 29 years old") has declarative or interrogative meaning (Beun, 1989). It also became clear that a separate category is required where a command is both Additive and Retracting at the same time, such as a *Modifying Command*. Furthermore, as displayed in Table 1 at 9:08, an action at one layer may not correspond to an action in another layer at all. Most interestingly, we discovered that knowledge authoring is full of what we call *Hidden Acts*, which do not correspond with actual Protégé actions, but which turn out to play an important role in the interaction, as evidenced by subjects' spoken utterances. These include, most prominently, various types of goal handling, such as the adoption, revision, satisfaction, and abandonment of goals. At the dialogue level, these correspond to the expression of desires (when a goal is set aside for later) and intentions (when a goal is starting to be pursued).

How does our annotation scheme compare to schemes for the analysis of natural language dialogue, such as the ISO-standard of (Bunt et al., 2010)? Similarities with our own scheme are easy to find: the distinction between information seeking and informative providing is an obvious example. Important differences exist as well. The ISO standard has been developed for annotating a wide class of dialogues. Our own scheme targets a specific kind of "asymmetric" dialogue, between a person and a machine. Goals can only be expressed by the person; *what-if* questions can only be posed by the person (the system responds, for example, with an *Expressive Statement*). Furthermore, the person cannot, at present, make

Expressive Statements (given that the addition of an axiom has been modelled as a *Command*). *Proposals* (as opposed to *Commands*) are only open to the system. Perhaps the most interesting differences between the two schemes relate to the expression of goals, which appears to be absent from the ISO standard. A few other natural-language related annotation schemes (e.g, (Leech and Weisser, 2003)) do allow the expression of "desires", though the distinction between desires and intentions – reminiscent of Bratman-style "Beliefs, Desires and Intentions" (Bratman, 1999) – is not made.

4 Where do we go from here?

Our annotation scheme has been updated to reflect the observations reported in the previous section, adding categories like *Expression of Intention* and *Expression of Desire*. Having offered tentative answers to the three questions mentioned at the end of the introduction to this paper, how do we plan to proceed?

First, we aim to further improve our understanding of knowledge authoring. Our experiment has given us some relevant insights, as incorporated in our modified annotation scheme, but the validity of these insights has yet to be proven. To do this we shall write an explicit annotation manual, and ask annotators to use these in a full-scale experiment. As before, the study proper will have as participants ontologists working on an ontology they either own or collaborate on in authoring and will not necessarily be computer scientists. Thus the study proper will observe participants in their everyday work, rather than a task that was merely an artefact of the study. In such a setting, there is the potential for a *think aloud* protocol to distort 'normal working practices', which was not an issue in the pilot study. After the experiment they will be asked to clarify their actions (retrospective think-aloud), and to elaborate on what they were trying to do (Olmsted-Hawala et al., 2010). Next, we shall ask annotators to use the annotation manual (and an annotation tool based on the categories defined in the manual) to annotate a number of episodes from the full-scale experiment. Inter-annotator agreement will then be measured using standard methods (for example, the κ test), which should tell us whether our annotation scheme, and the annotations that are based on it, are clear. Where necessary the scheme will be adapted in light of the experiment and the editors' clarifications.

As explained in the Introduction, we aim to exploit our improved understanding of knowledge authoring by designing an improved knowledge authoring interface. This will be a difficult step, which will start with an in-depth study of the outcomes of the full-scale experiment, trying to understand what ontology authors were attempting to do, and how their actions may be better supported by an improved knowledge authoring interface. It is too early to say with certainty what these improvements will be but, in light of our pilot studies, it seems plausible that the new interface will allow ontology authors to manipulate goals at different levels ("I'd like this concept to be subsumed by concept X", "I'd like this concept to have an infinite extension", "This goal has now been fulfilled and can therefore be deleted"), ask competency questions (such as questions that knowledge authors are often asked to write down before starting the authoring process proper) and ask *what-if* questions that allow users to explore the consequences of potential actions without as yet committing to them. Ultimately, we envisage implementing the new dialogue moves as a plug-in to Protégé-style knowledge authoring software.

References

Abgaz, Y., M. Javed, and C. Pahl (2012). Analyzing impacts of change operations in evolving ontologies. In *ISWC Workshops: Joint Workshop on Knowledge Evolution and Ontology Dynamics*.

Bernardi, R., D. Calvanese, and C. Thorne (2012). Designing efficient controlled languages for ontologies. In *Computating Meaning* (Vol 4 ed.). Springer.

Beun, R. (1989). *The recognition of declarative questions in information dialogues*. Ph. D. thesis, University of Tilburg.

Bratman, M. (1999). *Intention, plans, and practical reason*. Cambridge University Press.

Bunt, H. (1994). Context and dialogue control. *Think Quarterly 3*(1), 19–31.

Bunt, H., J. Alexandersson, J. Carletta, J.-W. Choe, A. C. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria, and D. Traum (2010). Towards an iso standard for dialogue act annotation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta. European Language Resources Association.

Buxton, W. (1983). Lexical and pragmatic considerations of input structures. *Computer Graphics 17*(1), 31–37.

Card, S., T. Moran, and A. Newell (1986). *The psychology of human-computer interaction*. L. Erlbaum Associates Inc.

Dzbor, M., E. Motta, J. M. Gomez, C. Buil, K. Dellschaft, O. Görlitz, and H. Lewen (2006). D4.1.1 analysis of user needs, behaviours & requirements wrt user interfaces for ontology engineering. Technical report, NeOn Project.

Foley, J. and A. Van Dam (1982). *Fundamentals of interactive computer graphics*, Volume 1. Addison Wesley Longman Publishing Co.

Fuchs, N., U. Schwertel, and R. Schwitter (1999). Attempto controlled english not just another logic specification language. In *Logic-Based Program Synthesis and Transformation*, Volume 1559, pp. 1–20. Springer.

Ginzburg, J. (1994). An update semantics for dialogue. In *Proceedings of the 1st International Workshop on Computational Semantics*.

Goncalves, R. S., B. Parsia, and U. Sattler (2011). Categorising logical differences between owl ontologies. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1541–1546. ACM.

Leech, G. and M. Weisser (2003). *Generic speech act annotation for task-oriented dialogues*, pp. 441–446. Centre for Computer Corpus Research on Language Technical Papers, Lancaster University.

Moran, T. (1981). The command language grammar: A representation for the user interface of interactive computer systems. *International Journal of Man-Machine Studies 15*(1), 3–50.

Nielsen, J. (1986). A virtual protocol model for computer-human interaction. *International Journal of Man-Machine Studies 24*(3), 301–312.

Olmsted-Hawala, E., E. Murphy, S. Hawala, and K. Ashenfelter (2010). Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In *the 28th international conference on Human factors in computing systems*, pp. 2381–2390.

Power, R. (2009). Towards a generation-based semantic web authoring tool. In *Proceedings of the 12th European Workshop on Natural Language Generation*, Stroudsburg, PA, USA, pp. 9–15. Association for Computational Linguistics.

Rector, A., N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens, H. Wang, and C. Wroe (2004). Owl pizzas: Practical experience of teaching owl-dl: Common errors & common patterns. In *Engineering Knowledge in the Age of the Semantic Web*, Volume 3257, pp. 63–81. Springer.

Schwitter, R. (2002). English as a formal specification language. In *Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on*, pp. 228 – 232.

Sowa, J. (2004). Common logic controlled english. `http://www.jfsowa.com/clce/specs.htm`. Accessed: 27/11/2012.

Towards a Tight Integration of Syntactic Parsing with Semantic Disambiguation by means of Declarative Programming*

Yuliya Lierler
University of Nebraska at Omaha
ylierler@unomaha.edu

Peter Schüller
Sabancı University
peterschueller@sabanciuniv.edu

Abstract

We propose and advocate the use of an advanced declarative programming paradigm – answer set programming – as a uniform platform for integrated approach towards syntax-semantic processing in natural language. We illustrate that (a) the parsing technology based on answer set programming implementation reaches performance sufficient for being a useful NLP tool, and (b) the proposed method for incorporating semantic information from FRAMENET into syntactic parsing may prove to be useful in allowing semantic-based disambiguation of syntactic structures.

1 Introduction

Typical natural language processing (NLP) system consists of at least several components including syntactic and semantic analyzers. A common assumption in the design of an NLP system is that these components are separate and independent. On one hand, this allows researchers an abstraction necessary to promote substantial steps forward in each task, plus such a separation permits for more convenient, modular software development. On the other hand, constraints from "higher level" processes are frequently needed to disambiguate "lower level" processes. For example, consider the syntactically ambiguous sentence

$$\textit{I eat spaghetti with chopsticks.} \tag{1}$$

Its verb phrase allows for two syntactic structures:

$$\dfrac{\dfrac{\dfrac{eat}{(VP/PP)/NP} \quad \dfrac{spaghetti}{NP}}{VP/PP} \qquad \dfrac{with\ chopsticks}{PP}}{VP} \qquad \dfrac{\dfrac{eat}{VP/NP} \quad \dfrac{\dfrac{spaghetti}{NP} \quad \dfrac{with\ chopsticks}{NP\backslash NP}}{NP}}{VP} \tag{2}$$

In the former, the prepositional phrase "$with\ chopsticks$" modifies the verbal phrase "$eat\ spaghetti$", and in the latter, it modifies the noun phrase "$spaghetti$". The sentence

$$\textit{I eat spaghetti with meatballs} \tag{3}$$

is syntactically ambiguous in a similar manner. In order to assign the proper syntactic structure to each of these sentences one has to take into account *selectional restrictions*, i.e., the semantic restrictions that a word imposes on the environment in which it occurs. For instance, in (1) the fact that a chopstick is an instrument suggests that "$with\ chopsticks$" modifies "$eat\ spaghetti$" as a tool for eating. Thus, an approach that integrates syntactic and semantic processing is essential for proper analysis of such sentences. Modern statistical methods, dominant in the field of syntactic analysis, take into account

[*] We would like to thank Vladimir Lifschitz and Ray Mooney for useful discussions and comments related to the topic of this work. Peter Schüller was supported by a TUBITAK scholarship.

selectional restrictions *implicitly* by assigning most probable syntactic structure based on observed co-occurrences of words and structures in corpora. Yet, this is often not sufficient. Sentences (1) and (3) illustrate this point, as the advanced parsers, including Stanford and Berkeley systems, do not produce proper syntactic representations for these sentences: instead they favor the same structure for both of them.[1] Similarly, semantic role labelers (joint syntactic-semantic parsers) such as SEMAFOR (Das et al., 2010) and LTH (Johansson and Nugues, 2007) display the same issue. The FRAMENET project (Baker et al., 1998) provides information that can disambiguate sentences (1) and (3). For instance, the frame *food* corresponds to the word "*spaghetti*". This frame contains information that *food* only takes other *food* as constituents. Thus modifying "*spaghetti*" with "*chopsticks*" in a parse tree for (1) yields a forbidden situation.

In this paper we present preliminary work on a system for natural language parsing that targets a tight integration of syntax and semantics. We illustrate its ability to take into account both quantitative and qualitative data available for processing natural language, where the former stems from statistical information available for natural language and the latter stems from lexical and commonsense knowledge available in lexical datasets such as FRAMENET.

Lierler and Schüller (2012) developed a Combinatory Categorial Grammar (CCG) parser ASPCCGTK[2]. A distinguishing feature of ASPCCGTK is that its design allows for synergy of both quantitative and qualitative information. First, it relies on the C&C part-of-speech *supertagger* (Clark and Curran, 2007) – built using latest statistical and machine learning advances in NLP. Second, its implementation is based on a prominent knowledge representation and reasoning formalism — answer set programming (ASP), see Brewka et al. (2011). ASP constitutes a convenient framework for representing constraints posed by selectional restrictions *explicitly*; thus we can augment implicit information available from statistical part-of-speech tagging with qualitative reasoning. We believe that the ASPCCGTK parser is a strong ground for designing a systematic, elaboration tolerant, knowledge intensive approach towards an integrated syntax-semantics analysis tool. Performance results on ASPCCGTK reported in (Lierler and Schüller, 2012) suggest that the "planning" approach adopted for parsing in the system scales to sentences of length up to 15 words. It may be sufficient for a number of applications: for example, 6.87 is the average number of words in sentences in the GEOQUERY corpus (Zelle and Mooney, 1996). But, in order for ASPCCGTK to become a viable NLP technology it is important to address the issue of its scalability.

The two contributions of this paper are as follows. First we demonstrate how use of the Cocke-Younger-Kasami (CYK) algorithm (Kasami, 1965) enhances the performance of ASPCCGTK. We evaluate the new approach implemented in ASPCCGTK on the CCGbank corpus (Hockenmaier and Steedman, 2007) and report the results. Second we propose and illustrate the method on how (a) FRAMENET can be used for properly disambiguating sentences (1) and (3), and (b) how this information is incorporated into the ASPCCGTK system. As a result we are able to use the ASPCCGTK parser to generate only the expected syntactic structures for the sentences in question.

In the future, we will automate a process of extracting selection restriction constraints from the data available in FRAMENET, by building an interface between ASPCCGTK and FRAMENET. CCGbank will provide us with extensive real world data for evaluating our approach. Once successful, we will look into expanding the approach to the use of other semantic annotations datasets for lexical items such as VERBNET, PROPBANK, NOMBANK and others for more complete sets of lexical constraints.

2 Extending ASPCCGTK for parsing CCG with CYK in ASP

Combinatory Categorial Grammar (Steedman, 2000) is a formalism that uses a small set of combinatory rules and a rich set of categories. Categories are either atomic such as NP, or complex such as $S\backslash NP$, which is a category for English intransitive verbs. The category $S\backslash NP$ states that an NP to the left of the

[1] Thanks to Nathan Schulte for carrying out the research supporting this claim by investigating the behavior of nine state-of-the-art parsers listed at `http://nlp.stanford.edu/software/stanford-dependencies.shtml`.

[2] `http://www.kr.tuwien.ac.at/staff/former_staff/ps/aspccgtk`

word will result in a sentence S. Given a sentence and a lexicon containing a set of word-category pairs, we replace words by appropriate categories and then apply combinators. For example, in the former derivation in (2), "*eat*" has category $(VP/PP)/NP$ and "*spaghetti*" has category NP. The combinator used in this derivation is *forward application* (*fa*)

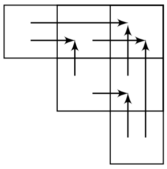
$$\frac{A/B \quad B}{A}\,fa \tag{4}$$

where A and B are variables that can be substituted for CCG categories. Applying forward application to "*eat*" and "*spaghetti*" substitutes A with VP/PP and B with NP and yields VP/PP. An input sentence is part of a grammar if some sequence of applying combinators results in the category S at the root of the parse tree.

The implementation of ASPCCGTK is based on answer set programming – a declarative logic programming paradigm. ASP roots in answer set semantics of logic programs (Gelfond and Lifschitz, 1988). The idea of ASP is to represent a problem by a program whose answer sets correspond to solutions. For example, for parsing we encode the grammar and the input sentence in a way that each answer set corresponds to a valid parse tree of the input. Unlike in an imperative style of programming, in declarative programming we describe a specification of the problem, which expresses what the program should accomplish rather than prescribing how to do it. Answer set solvers use this specification to efficiently navigate through a search space and find solutions to the problem. For a more detailed and yet brief introduction of CCG and ASP we refer the reader to (Lierler and Schüller, 2012).

The CYK (Cocke, Younger, and Kasami) algorithm for context-free-grammars was initially published by Kasami (1965). It can be extended to CCG using ideas from Lange and Leiß (2009). Given an input of n words, CYK operates on an $n \times n$ triangular chart. Words in the input are associated with categories in the diagonal of the chart. Combinatory rules combine categories from two chart cells into a single category in another "corresponding" cell. We illustrate these intuitions using a 3×3 chart:

$$\begin{tikzcd} A/B \arrow[r] & A \\ & B \arrow[u] \end{tikzcd}$$



$$\begin{tikzcd} \texttt{Pleft,Pj} \arrow[r] & \texttt{Pi,Pj} \\ & \texttt{Pi,Pdown} \arrow[u] \end{tikzcd}$$

1,1	2,1	3,1
	2,2	3,2
		3,3

(5)

An input is recognized as part of the grammar if the top right chart cell contains the category `S` after successive application of combinators to the chart. A realization of CYK for recognition of context-free grammars in ASP was described by Drescher and Walsh (2011). First, we adapt their approach to CCG. Second, we extend it to the task of generating parse trees as we are not only interested in recognizing grammatical inputs but also in producing appropriate parses.

We now show parts of our realization of CYK in the ASP formalism. We represent a chart using a predicate `grid(Pi,Pj,Cat)` and initialize the diagonal using the rule

```
grid(P,P,C) :- category_at(C,P).
```

where the `category_at` predicate is obtained from tagging the input with the C&C supertagger.

Forward application is realized as follows (grid variables are shown in (5), X and Y are variables that stand for CCG categories):

```
applicable(fa,Pj,Pi,Pleft,Pdown,X,rfunc(X,Y),Y) :-
  grid(Pleft,Pi,rfunc(X,Y)), grid(Pj,Pdown,Y).
grid(Pj,Pi,X) :- applicable(_,Pj,Pi,_,_,X,_,_).
```

where `rfunc(X,Y)` encodes complex category of the form X/Y. The first rule defines where the *fa* combinator can be applied to within the chart. The second rule defines which categories this application creates. For obtaining parse trees, we "guess" for each instance of `applicable` if that combinator is actually applied (`SrcLeft`, `SrcDown`, and `Result` stand for CCG categories in the CYK grid):

```
{ applied(Comb,Pj,Pi,Pleft,Pdown,Result,SrcLeft,SrcDown) } :-
  applicable(Comb,Pj,Pi,Pleft,Pdown,Result,SrcLeft,SrcDown).
```

The curly bracket construct in the head of this rule is what expresses the guess as we can intuitively read this rule as follows: an expression in the head *may* hold in case if the body of the rule holds.

To obtain only valid parse trees, we furthermore (i) add rules that constraint the selection of multiple `applied` combinators in one cell, (ii) define reachability of diagonal chart cells from the S category in the top right cell, and (iii) add rules that require all diagonal cells to be reachable.

We believe that the possibility of explicitly representing alternatives and then restricting them by expressing appropriate conditions using declarative means makes ASP a powerful tool for working with ambiguities in the field of NLP.

We conducted empirical experiments to compare the performance of the original ASPCCGTK and the ASPCCGTK enhanced with CYK as described here. We report average times and number of timeouts when parsing all sentences of Section 00 of the CCGbank corpus (Hockenmaier and Steedman, 2007) using a timeout of 1200 seconds. The sentences were chunked and tagged by the C&C supertagger. The benchmark results show that the CYK approach has a significant performance advantage over the old approach. Columns show average time in seconds for groups of sentences of a certain length. Number in parenthesis represents the number of timeouts.

Number of Words	1-10	11-15	16-20	21-25	26-30	31-35	36-40	41+
Number of Sentences	195	285	345	330	287	224	118	129
ASPCCGTK (CYK)	0.04	0.18	0.49	1.04	2.11	3.21	6.66	27.01(3)
ASPCCGTK (original)	0.13	1.07	4.90	20.93	68.21(1)	194.59(2)	342.88(24)	497.93(75)

3 Semantic Disambiguation using FRAMENET

FRAMENET is a dataset of semantic relations based on Frame Semantics (Fillmore and Baker, 2001). Lexical items *evoke* certain *frames* that contain *frame elements*; for example, "*eat*" evokes an *ingestion* frame and everything that is of *semantic type* **food** evokes a *food* frame. Sample information available in the *ingestion* and *food* frames follow:

Frame	*Frame Element*	*Semantic Type*
ingestion	INGESTOR	**sentient**
	INGESTIBLE	**ingestible**
	INSTRUMENT	**tool**
	MANNER	**manner**

Frame	*Frame Element*	*Semantic Type*
food	CONSTITUENT	**food_constituent**

Each frame element is a slot that *may* be filled only by elements of the correct semantic type. Types are organized in a taxonomy. For instance, the following part of the taxonomy is relevant to this presentation:

tool is_a **instrument** **food** is_a **ingestible** **food** is_a **food_constituent**.

We propose a concept of a "semantically coherent" parse tree. Information from FRAMENET allows us to disambiguate semantically coherent and incoherent trees. We now make these ideas precise. Each node in a tree is annotated with a *tag* – either a distinguished tag $\perp$ or a pair $T||F$ where both T and F are sets consisting of semantic types. Each leaf of a tree is assigned a tag $T||F$ in accordance with FRAMENET information for a corresponding word. The set T contains the semantic types associated with the leaf-word. For instance, for word "*spaghetti*", this set T_{sp} is $\{$**food**, **food_constituent**, **ingestible**$\}$. The set F contains the semantic types associated with the frame elements of a frame evoked by a leaf-word. For instance, for word "*eat*" that evokes the *ingestion* frame, this set F_{eat} is

$\{$**sentient**, **ingestible**, **tool**, **manner**$\}$.

To define a tag for a non-leaf node of a tree we introduce the following terminology. Any non-leaf node in a CCG parse tree is a parent of two children: a *functor* and an *argument*. Depending on semantic information assigned to nodes, they act as functors or arguments. For a non-leaf node p, we define a tag $T_p||F_p$ as follows

$$T_p||F_p = \begin{cases} \bot & \text{if a tag of either } f \text{ or } a \text{ is } \bot \\ \bot & \text{if } F_f \cap T_a = \emptyset \\ T_f||(F_f \setminus \{s\}) & \text{if there is a semantic type } s \in F_f \cap T_a \end{cases}$$

where f and a stand for a functor and an argument children of p, respectively. Pairs $F_f||T_f$ and $F_a||T_a$ correspond to tags of these children. We say that a parse tree is *semantically coherent* if there is no node in the tree annotated by the $\bot$ tag.

Recall the syntactic structures (2) corresponding to the verb phrase of sentence (1). The annotated counterpart of the former structure follows[3]:

$$\dfrac{\dfrac{\dfrac{\mathit{eat}}{(VP/PP)/NP : \emptyset||F_{eat}} \quad \dfrac{\mathit{spaghetti}}{NP : T_{sp}||\{\textbf{\textsf{food\_constituent}}\}}}{VP/PP : \emptyset||\{\textbf{\textsf{sentient}},\textbf{\textsf{tool}},\textbf{\textsf{manner}}\}} \quad \dfrac{\mathit{with\ chopsticks}}{PP : \{\textbf{\textsf{tool}},\textbf{\textsf{instrument}}\}||\emptyset}}{VP : \emptyset||\{\textbf{\textsf{sentient}},\textbf{\textsf{manner}}\}}$$

This subtree is semantically coherent. On the other hand, part of the later structure in (2) constitutes semantically incoherent subtree:

$$\dfrac{\dfrac{\mathit{spaghetti}}{NP : T_{sp}||\{\textbf{\textsf{food\_constituent}}\}} \quad \dfrac{\mathit{with\ chopsticks}}{NP\backslash NP : \{\textbf{\textsf{tool}},\textbf{\textsf{instrument}}\}||\emptyset}}{NP : \bot}$$

To implement described process within ASPCCGTK approach, we first manually specify a dictionary that contains FRAMENET information sufficient for annotating leaf nodes stemming from the words in an input sentence. We then use logic rules to (a) define annotations for non-leaf nodes of parse trees and (b) restrict the produced parse trees only to these that are semantically coherent. On the sentences (1) and (3), the ASPCCGTK parser implementing this approach is capable to enumerate only and all semantically coherent parses that correspond to syntactic structures expected for the sentences.

4 Conclusions and Future Work

In this work we propose and advocate the use of advanced declarative programming paradigm – answer set programming – as a uniform platform for integrated approach towards syntax-semantic processing in NLP. We illustrate that the CCG parser ASPCCGTK based on an ASP implementation reaches performance sufficient for being a useful NLP technology by taking advantage of the data structures of the CYK algorithm. Even though ASP has a high worst-case complexity, a related declarative paradigm with the same worst-case complexity was shown to be effective for solving NLP problems: the usage of Integer Linear Programming in (Roth and Yih, 2007). We also propose a method for disambiguating syntactic parse trees using the semantic information stemming from the FRAMENET dataset and implement it within the ASPCCGTK parser. This implementation results in the first step towards a synergistic approach in syntax-semantic processing by means of technology such as ASP. There is an open question on how to automatically fetch relevant information from FRAMENET in order to make the proposed implementation widely usable. This is the subject of ongoing and future work. One reasonable direction to explore is assess the usability of FRAMENET-based semantic role labeling systems for our purposes, in particular, LTH and SEMAFOR. The CCGbank will serve us as a test bed for evaluating the effectively of proposed method and directing this research. The source code of the reported implementation is available online at the ASPCCGTK website under version 0.3.

[3] The definition of semantically coherent trees presented here is a simplified version of a more complex construct, which takes into account functors that carry no semantic type information by themselves (for example, a functor corresponding to a word "*with*") but rather inherit this information from its argument.

References

Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, San Mateo, CA, pp. 86–90.

Brewka, G., I. Niemelä, and M. Truszczyński (2011). Answer set programming at a glance. *Communications of the ACM 54(12)*, 92–103.

Clark, S. and J. R. Curran (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics 33*(4), 493–552.

Das, D., N. Schneider, D. Chen, and N. A. Smith (2010). Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, Stroudsburg, PA, USA, pp. 948–956. Association for Computational Linguistics.

Drescher, C. and T. Walsh (2011). Modelling grammar constraints with answer set programming. In *International Conference on Logic Programming*, Volume 11, pp. 28–39.

Fillmore, C. J. and C. F. Baker (2001). Frame semantics for text understanding. In *WordNet and Other Lexical Resources, NAACL Workshop*.

Gelfond, M. and V. Lifschitz (1988). The stable model semantics for logic programming. In R. Kowalski and K. Bowen (Eds.), *Proceedings of International Logic Programming Conference and Symposium (ICLP'88)*, pp. 1070–1080. MIT Press.

Hockenmaier, J. and M. Steedman (2007). CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguististics 33*(3), 355–396.

Johansson, R. and P. Nugues (2007). Lth: semantic structure extraction using nonprojective dependency trees. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, Stroudsburg, PA, USA, pp. 227–230. Association for Computational Linguistics.

Kasami, T. (1965). An efficient recognition and syntax analysis algorithm for context-free languages. Technical Report AFCRL-65-758, Air Force Cambridge Research Laboratory, Bedford, Massachusetts.

Lange, M. and H. Leiß (2009). To CNF or not to CNF? An efficient yet presentable version of the CYK algorithm. In *Informatica Didactica 8*. Universität Potsdam.

Lierler, Y. and P. Schüller (2012). Parsing combinatory categorial grammar via planning in answer set programming. In *Correct Reasoning*, Volume 7265 of *Lecture Notes in Computer Science*, pp. 436–453. Springer.

Roth, D. and W. Yih (2007). Global inference for entity and relation identification via a linear programming formulation. In L. Getoor and B. Taskar (Eds.), *Introduction to Statistical Relational Learning*. MIT Press.

Steedman, M. (2000). *The syntactic process*. London: MIT Press.

Zelle, J. M. and R. J. Mooney (1996). Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, Portland, OR, USA, pp. 1050–1055.

Situated Utterances and Discourse Relations

Matthew Stone[†§], Una Stojnic[‡§] and Ernest Lepore[‡§*]
[†]Computer Science, [‡]Philosophy, [§]Cognitive Science
Rutgers University
matthew.stone@rutgers.edu, ustojnic@eden.rutgers.edu, lepore@ruccs.rutgers.edu

Abstract

Utterances in situated activity are about the world. Theories and systems normally capture this by assuming references must be resolved to real-world entities in utterance understanding. We describe a number of puzzles and problems for this approach, and propose an alternative semantic representation using discourse relations that link utterances to the nonlinguistic context to capture the context-dependent interpretation of situated utterances. Our approach promises better empirical coverage and more straightforward system building. Substantiating these advantages is work in progress.

1 Introduction

People exhibit sophisticated strategies for using language to coordinate their ongoing practical activities (Clark and Krych, 2004). Applications such as human–robot interaction must also also support these strategies (Kruijff et al., 2012). A key question is how to track speakers' reference to the environment. Julia Child's (1), from the "omelette" episode of *The French Chef*, illustrates the issues:

(1) There's your omelette, forming itself in the bottom of the pan.

The accompanying video is a close-up of a stovetop from above, with the omelette Child is cooking front and center. To understand what Child is doing, we need to track the obvious connection between her utterance and *that omelette* in *that pan*.

In this paper, we explore a novel and surprising proposal about how do this: *discourse relations*. Discourse relations are kinds of speech acts, which can connect an utterance to ongoing conversation, establish the coherence of what is said, and allow inference to implicatures and other aspects of the speaker's mental state (Asher and Lascarides, 2003, 2013). Child's (1) we suggest, gets its coherence, in part, as a *report of what's visible in that situation*. Child's reference to the omelette and the pan is a consequence of this interpretation, and so need not and should not be separately represented.

Prior work focuses on symbols that refer directly to the world and placeholders that abstract reference; see Section 2. We introduce our alternative and contrast it with these approaches in Section 3. On the theoretical side, we argue in Section 4, our approach captures meaning more precisely and provides a better account of what's needed to understand and disambiguate situated utterances. On the practical side, we argue in Section 5, our approach provides a more tractable interface between linguistic processing, situated perception, and deliberation.

2 Background

Our work arises out of an interest in combining grounded representations of meaning with coherence approaches to discourse. To appreciate the issues, consider a classic case (Bolt, 1980). A user utters (2), while pointing first at an object in the environment and then at the place it should go.

(2) Put that there.

[*]This paper is based on work supported by NSF IIS-1017811. We acknowledge the helpful anonymous reviews.

Kaplan (1989) urges us to treat demonstratives in cases such as (2) as *directly referential*. In using them, the speaker gives information about *those very things*. Computationally, this means that the objects of demonstrations should be represented using symbols that are locked onto their referents system-wide. In multimodal interfaces, like Bolt's (1980), the targets are graphical objects and other locations on a computer display, for which appropriate internal representations are readily available. For robotics, the natural strategy is to represent referents using perceptually-grounded symbols anchored in the real world, as do Yu and Ballard (2004) for example.

Speakers *use* many utterances referentially, even if the utterances don't have referential semantics (Kripke, 1977). Systems need to recognize and represent the speaker's referential communicative intentions in those cases too (Allen and Perrault, 1980). Bolt's system, for example, responded to definite descriptions, as in (3), the same way as it did to demonstratives: it moved *those things*.

(3) Put the cruise ship north of the Dominican Republic.

When it's problematic to use grounded symbols, we can use *discourse referents* alongside them. Formally, a discourse referent is just a free variable, but it can be associated with *anchoring constraints* that describe how it is supposed to be linked up with the world (Zeevat, 1999). In practice, anchoring involves representing interpretation in two separate tiers (Luperfoy, 1992). Meaning is represented via variables, the world is represented via suitably grounded symbols, and an evolving assignment of values to variables embodies the system's understanding of the real-world reference of expressions in discourse.

Reconciling grounded reference and discourse anaphora is the job of discourse semantics. We can see this in the planning dialogue of (4), for example.

(4) a. A: Let's put the cruise ship south of Florida.
b. B: That won't fit there.

We need to represent A's utterance as a proposal—a specific kind of problem-solving move that advocates a particular course of action and commits the speaker to following through on it if others concur. The move focuses attention on the entities involved in carrying it out: here, the ship, Florida and the region to its south. Meanwhile, B's response in (4b) is a rejection—a move that offers a negative assessment of the current proposal and commits the speaker against adopting it. Note that B's references with *that* and *there* succeed even if B produces the utterance without any accompanying gesture or demonstration, because the referent has been activated by an earlier mention in a related utterance (Gundel et al., 1993).

The simplest approach to discourse organization is to represent the state of the discourse with an information state (Poesio and Traum, 1997) and associate each move with an appropriate update. For example, we can model utterances such as (2) and (3) as making moves that contribute step-by-step to broader problem-solving activity; see Lochbaum (1998) or Blaylock (2005). Normally, the information state specifies once and for all how each thread of ongoing activity places relevant real-world entities at the center of attention (Grosz and Sidner, 1986; Poesio and Traum, 1997).

We advocate a different approach, based on *discourse coherence* (Kehler, 2002; Asher and Lascarides, 2003). The idea is that discourse is fundamentally composed of *relational* contributions, which establish connections that link each utterance by inference to segments of the preceding conversation. The interpretation of an utterance therefore implicitly refers to the interpretation of some prior discourse and comments on it. On coherence approaches, how an utterance *attaches* to the discourse determines what entities are prominent in interpreting it (Hobbs, 1979). Coherence theory does *not* naturally characterize discourse in terms of state-by-state updates to an overarching model of information and attention. Kehler and colleagues' (5) illustrates what's at stake (Kehler et al., 2008).

(5) Phil tickled Stanley, and Liz poked him.

When we understand the second clause of (5) as a description of a parallel to the first, we prefer to resolve *him* to Stanley. When we understand it to describe its results, we prefer to resolve *him* to Phil. For coherence theory, the two interpretations of the second clause *relate* it to the first, and the relation is what suggests prominent resolutions for its references. The relation in turn structures the discourse into higher level units that shape possibilities for attaching subsequent utterances.

3 Our Proposal

Coherence theories start from the observation that understanding utterances involves recognizing the implicit relationships that connect ideas together. The same is true, we argue, for utterances like (1). Child is not just giving the next step in making an omelette, or giving her audience new information about the principles of cooking. She's describing what's happening on the screen, in terms she expects her audience to confirm for themselves by examining what they see. An interpreter who doesn't recognize this about (1) has failed to understand it.

To sketch the key formal ingredients of our account, we use a simple dynamic semantics (Muskens, 1996) and an expressive ontology of situations and eventualities (Hobbs, 1985; Kratzer, 2002). Dynamic semantics represents meanings as sequences of updates $[\mathbf{v}|\varphi]$ that introduce discourse referents $\mathbf{v}$ and characterize them with conditions φ. We add an update $\langle \pi x c \rangle$ introducing a discourse referent x perceptually grounded in c, and an update $\langle \sigma x s \rangle$ introducing x as the central entity in grounded situation s (provided s does uniquely distinguish one). Where necessary, ∂K marks update K as presupposed.

Situations are parts of the world, capturing particular states of particular objects, perhaps as located in particular spatial regions and changing over particular temporal intervals. Propositions are true in situations; this is important for perceptual, causal and default reasoning. Eventualities, including Davidsonian events, turn out to be situations that exemplify propositions in characteristic ways. We capture discourse coherence by specifying relations among situations; these can be discourse referents for situations introduced by utterances or grounded references to parts of the speech situation.[1]

We capture the interpretation of (2) by specifying the dynamics of discourse referents and their grounded interpretations as in (6), using c_1 for *that* and c_2 for *there*.

(6) $\langle \pi x c_1 \rangle ; \langle \pi y c_2 \rangle ; [e | \mathit{command}(e), \mathit{put}(e, x, y)]$

Things get more interesting when we factor in coherence. We formalize (1) as in (7).

(7) $[e|\mathit{Summary}(s_0, e)]; \partial[o, p|\mathit{omelette}(o), \mathit{pan}(p)]; [|\mathit{forming\text{-}self\text{-}in}(e, o, p)]$

The discourse relation $\mathit{Summary}(s_0, e)$ captures the interpretive connection between the utterance describing e and what's happening simultaneously on the screen in situation s_0. Like all coherence relations, *Summary* reflects semantic and pragmatic constraints. Semantically, e must be part of s_0. Following Kratzer (2002), this entails that the information describing e is true in s_0. Pragmatically, $\mathit{Summary}(s_0, e)$ holds only if the information describing e provides a good answer about "what's happening" in s_0. A summary appeals to broad, basic categories to provide essential information. We have in mind something like the "vital nuggets of information" needed to answer definition questions (Voorhees, 2003).

For "That's an omelette" we offer (8), which defines the central entity in situation s_0 as an omelette:

(8) $[e|\mathit{Summary}(s_0, e)]; \langle \sigma o s_0 \rangle; [|\mathit{omelette}(e, o)]$

The update $\langle \sigma o s_0 \rangle$ formalizes how the *discourse relation* makes entities prominent for reference, as we observed in (5). Such updates can capture the interpretation of demonstratives when there's no explicit pointing or demonstration in the utterance.

Not all situated utterances offer a *Summary* of an unfolding situation. For example, utterances can offer *Assessments* that invite the audience not to define what's happening but to appraise it. Take "Yummy!" In commenting on the food this way, the speaker expects the audience to join in her appreciation. A formal characterization of *Assessment* would appeal to the semantics of predicates of personal taste and the distinctive pragmatic functions of such judgments, perhaps following Crespo and Fernández (2011). And speakers can also link up questions and instructions to ongoing activity by suitable relations.

Summary and *Assessment* could also be used to formalize the interpretation of successive utterances by relating two described situations. In fact, utterances can relate both to ongoing activity and to previous

[1] Thus we use situations to capture discourse meaning, *not* to formalize events of speech or the common ground as in Poesio and Traum (1997). An alternative approach would follow Zeevat (1999) and Asher and Lascarides (2003) and use labels for DRSs to capture perceptual and discourse content in discourse relations.

discourse. For example, consider (9) and (10), taken from Vi Hart's origami proof of the Pythagorean theorem—a visual narrative much like Child's where utterances describe ongoing events on the screen.[2]

(9) We're just taking advantage of the symmetries of the square for the next step.

(10) This is where you're choosing how long and pointy or short and fat the right triangle is.

Hart uses (9) while folding a square into eight identical segments to explain how to do the folds. Hart uses (10) as she describes the next step of folding, to highlight its result for the proof. Thus, these utterances are linked to the accompanying activity but do not just report what's going on; and they're linked to the ongoing discourse as well. In fact, coherence theory already allows that utterances can bear multiple connections to prior discourse (Asher and Lascarides, 2003). The closest parallel may be that of multimodal communication, where Lascarides and Stone (2009) argue that utterances bear discourse relations both to prior utterances and to simultaneous gesture.

4 Empirical Adequacy

Combining grounded representations with discourse relations, specifically as in (8), makes it possible to give a better characterization of the logical form of demonstrative utterances in otherwise problematic cases. In particular, it captures how speakers and interpreters can rely on the world to disambiguate what they say and to understand one another.

Here's a telling case. It's the beginning of spring, 2012, and Jupiter and Venus are shining brightly very close together—just a few degrees apart—in the evening sky. The speaker has deployed a telescope facing a window over the western sky. When a visitor arrives, the speaker adjusts the telescope, then says, without any further demonstration, either (11) or (12).

(11) That's Jupiter. You can even see four moons.

(12) That's Venus. You can see the crescent.

We (and our informants) find these utterances unproblematic. But the coherence theory is required to get their interpretations right. These are comments on what's visible through the telescope. You can't see four of Jupiter's moons or the crescent of Venus with the naked eye and the speaker isn't suggesting otherwise. Moreover, the coherence relation is what's making it possible for the speaker to refer to Jupiter or Venus as *that*. To comment on the view through the telescope is to evoke whatever entity is centrally imaged in the telescope as a prominent candidate for reference. And nothing else will do. Given the astronomical conjunction, the speaker couldn't distinguish Jupiter from Venus by pointing, nor could the visitor judge which body the telescope was pointed at by the direction of the tube. Letting s_1 name the view through the telescope, we can formalize the key bits of interpretation:

(13) $[e|Summary(s_1,e)];\langle\sigma x s_1\rangle;[|jupiter(e,x)]$

(14) $[e|Summary(s_1,e)];\langle\sigma x s_1\rangle;[|venus(e,x)]$

The representations get the meaning right. More importantly, they explain how the visitor can recover the logical form and understand the speaker's point by recognizing the relationship that makes the speaker's utterance coherent, even though the visitor can't identify which specific body the speaker is referring to until the visitor looks through the telescope for herself. By contrast, if all you had was representations like (6), grounded representations of deixis that made reference explicit, you'd incorrectly predict that there's an ambiguity to resolve in (11) and (12) even *after* you understand them as comments about the view through the scope. You'd have two grounded symbols for bright objects in the western sky, and you'd have to *pick* one as the referent of the speaker's demonstration—or ask for clarification. We take this as strong evidence against the idea that speakers and hearers must coordinate directly on demonstrative referents, a common view in both formal and computational semantics (Neale, 2004; Stone, 2004).

[2] `http://www.youtube.com/watch?v=z6lL83wl31E`

5 Cognitive Architecture

Systems with grounded language interpretation have to integrate language and perception. Real robotic understanding systems use a complicated inference process to resolve reference in situated utterances, in order to figure out what thing in the world is a reference of a demonstration, for example (Kruijff et al., 2012). It's not just a matter of selecting the right referent from a set of salient candidates based on linguistic constraints. Like people, robots have cameras with limited fields of view that must be pointed at what they see. So when you do a demonstration for a robot, it has to track the pointing movement to find the possible targets, much as a person would. There is substantial problem solving involved.

The same is true of many other grounded inference tasks involving domain reasoning. We may need the results to calculate the implications of what we hear, or even to select the most likely interpretation. But it's normally prohibitive to try to interleave that problem solving with fine-grained disambiguation, because it requires systems to solve these hard reasoning problems not just for the intended interpretation but for any candidate interpretation it considers.

Representing context dependence via coherence provides an attractive framework to divide interpretation into stages and minimize the problem solving that's necessary to compute logical form. Take (11) and its representation in (13). Here is a formalism that captures the meaning of the utterance while spelling out the further work that will be required to resolve reference. According to (13), when you look through the telescope, you'll find out what the referent of *that* is. Recognizing the logical form of the utterance this way should suffice for understanding. We expect that the discourse relation could be resolved based on shallow constraints on what information counts as a summary. And the system only needs to be tracking the ongoing activity enough to link utterances to relevant situations. It can ground its provisional understanding in perception and action as needed separately.

6 Conclusion and Future Work

We have considered grounded interpretations in coherent discourse, and argued that referential interpretations in cases like (1), (2) and (11) are understood and derived relationally. This requires representations of interpretation that explicitly link discourse entities with grounded symbols and track the heterogeneous prominence that these entities get in virtue of the diverse relationships that utterances can bear to ongoing activity. In brief: we relate our talk to the world around us through suitable discourse relations.

Our approach commits us to representing utterances with specific kinds of interpretive connections to the world. Our characterization of these connections is obviously provisional, and corpus and modeling work is necessary to flesh out the parameters of the approach. We have also suggested that these representations will be useful in designing systems that communicate about the environment where they can perceive and act. Experiments with prototypes are clearly necessary to substantiate this claim.

References

Allen, J. F. and C. R. Perrault (1980). Analyzing intention in utterances. *AIJ 15*(3), 143–178.

Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Asher, N. and A. Lascarides (2013). Strategic conversation. *Semantics and Pragmatics*.

Blaylock, N. J. (2005). *Towards Tractable Agent-Based Dialogue*. Ph.D. dissertation, Rochester.

Bolt, R. (1980). Put-that-there: Voice and gesture at the graphics interface. *Computer Graphics 14*(3), 262–270.

Clark, H. H. and M. A. Krych (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language 50*, 62–81.

Crespo, I. and R. Fernández (2011). Expressing taste in dialogue. In *SEMDIAL 2011: Proceedings of the 15th Workshop on the Seantics and Pragmatics of Dialogue*, pp. 84–93.

Grosz, B. J. and C. L. Sidner (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics 12*(3), 175–204.

Gundel, J. K., N. Hedberg, and R. Zacharski (1993). Cognitive status and the form of referring expressions in discourse. *Language 69*(2), 274–307.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science 3*(1), 67–90.

Hobbs, J. R. (1985). Ontological promiscuity. In *Proceedings of ACL*, pp. 61–69.

Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry, and H. Wettstein (Eds.), *Themes from Kaplan*, pp. 481–563. Oxford: Oxford University Press.

Kehler, A. (2002). *Coherence, Reference and the Theory of Grammar*. Stanford: CSLI.

Kehler, A., L. Kertz, H. Rohde, and J. L. Elman (2008). Coherence and coreference revisited. *Journal of Semantics 25*(1), 1–44.

Kratzer, A. (2002). Facts: Particulars or information units? *Linguistics & Philosophy 25*(5–6), 655–670.

Kripke, S. (1977). Speaker's reference and semantic reference. In P. A. French, T. Uehling, Jr., and H. K. Wettstein (Eds.), *Midwest Studies in Philosophy, Volume II*, pp. 255–276. Minneapolis: University of Minnesota Press.

Kruijff, G.-J., M. Janicek, and H. Zender (2012). Situated communication for joint activity in human-robot teams. *IEEE Intelligent Systems 27*(2), 27–35.

Lascarides, A. and M. Stone (2009). Discourse coherence and gesture interpretation. *Gesture 9*(2), 147–180.

Lochbaum, K. E. (1998). A collaborative planning model of intentional structure. *Computational Linguistics 24*(4), 525–572.

Luperfoy, S. (1992). The representation of multimodal user interface dialogues using discourse pegs. In *Proceedings of ACL*, pp. 22–31.

Muskens, R. (1996). Combining Montague semantics and discourse representation. *Linguistics & Philosophy 19*(2), 143–186.

Neale, S. (2004). This, that, and the other. In A. Bezuidenhout and M. Reimer (Eds.), *Descriptions and Beyond*, pp. 68–181. Oxford: Oxford University Press.

Poesio, M. and D. R. Traum (1997). Conversational actions and discourse situations. *Computational Intelligence 13*(3), 309–347.

Stone, M. (2004). Intention, interpretation and the computational structure of language. *Cognitive Science 28*(5), 781–809.

Voorhees, E. M. (2003). Evaluating answers to definition questions. In *Companion Volume of the Proceedings of HLT-NAACL – Short Papers*, pp. 109–111.

Yu, C. and D. H. Ballard (2004). On the integration of grounding language and learning objects. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI)*, pp. 488–494.

Zeevat, H. (1999). Demonstratives in discourse. *Journal of Semantics 16*(4), 279–313.

Gamification for Word Sense Labeling

Noortje J. Venhuizen
University of Groningen
n.j.venhuizen@rug.nl

Valerio Basile
University of Groningen
v.basile@rug.nl

Kilian Evang
University of Groningen
k.evang@rug.nl

Johan Bos
University of Groningen
johan.bos@rug.nl

Abstract

Obtaining gold standard data for word sense disambiguation is important but costly. We show how it can be done using a "Game with a Purpose" (GWAP) called *Wordrobe*. This game consists of a large set of multiple-choice questions on word senses generated from the Groningen Meaning Bank. The players need to answer these questions, scoring points depending on the agreement with fellow players. The working assumption is that the right sense for a word can be determined by the answers given by the players. To evaluate our method, we gold-standard tagged a portion of the data that was also used in the GWAP. A comparison yielded promising results, ranging from a precision of 0.88 and recall of 0.83 for relative majority agreement, to a precision of 0.98 and recall of 0.35 for questions that were answered unanimously.

1 Introduction

One of the core aspects of semantic annotation is determining the correct sense of each content word from a set of possible senses. In NLP-related research, many models for disambiguating word senses have been proposed. Such models have been evaluated through various public evaluation campaigns, most notably SenseEval (now called SemEval), an international word sense disambiguation competition held already six times since its start in 1998 (Kilgarriff and Rosenzweig, 2000).

All disambiguation models rely on gold standard data from human annotators, but this data is time-consuming and expensive to obtain. In the context of constructing the Groningen Meaning Bank (GMB, Basile et al., 2012), a large semantically annotated corpus, we address this problem by making use of crowdsourcing. The idea of crowdsourcing is that some tasks that are difficult to solve for computers but easy for humans may be outsourced to a number of people across the globe. One of the prime crowdsourcing platforms is Amazon's Mechanical Turk, where workers get paid small amounts to complete small tasks. Mechanical Turk has already been successfully applied for the purpose of word sense disambiguation and clustering (see, e.g., Akkaya et al., 2010; Rumshisky et al., 2012). Another crowdsourcing technique, "Game with a Purpose" (GWAP), rewards contributors with entertainment rather than money. GWAPs challenge players to score high on specifically designed tasks, thereby contributing their knowledge. GWAPs were successfully pioneered in NLP by initiatives such as 'Phrase Detectives' for anaphora resolution (Chamberlain et al., 2008) and 'JeuxDeMots' for term relations (Artignan et al., 2009). We have developed an online GWAP platform for semantic annotation, called Wordrobe. In this paper we present the design and the first results of using Wordrobe for the task of word sense disambiguation.

2 Method

Wordrobe[1] is a collection of games with a purpose, each targeting a specific level of linguistic annotation. Current games include part-of-speech tagging, named entity tagging, co-reference resolution and

[1] http://www.wordrobe.org/

word sense disambiguation. The game used for word sense disambiguation is called *Senses*. Below we describe the design of Wordrobe and the data used for *Senses*.

2.1 Design of Wordrobe

Wordrobe is designed to be used by non-experts, who can use their intuitions about language to annotate linguistic phenomena, without being discouraged by technical linguistic terminology. Therefore, the games include as little instructions as possible. All games share the same structure: a multiple-choice question with a small piece of text (generally one or two sentences) in which one or more words are highlighted, depending on the type of game. For each question, players can select an answer or use the skip-button to go to the next question.

In order to encourage players to answer a lot of questions and to give good answers, they are rewarded in two ways: they can collect *drawers* and *points*. A drawer is simply a unit of a few questions – the more difficult the game, the fewer questions are in one drawer. By completing many drawers, players unlock achievements that decorate their profile page. While drawers are used to stimulate answering many questions, points are used to motivate players to play with attention. The points are calculated on the basis of two factors: the *agreement* with other players who answered the same question and the *bet* that the player put at stake. Players can place a bet reflecting the certainty about their answer. The bet is always between 10% and 100% of the points that a question is worth. The default choice is a bet of 10% and once a player adjusts the bet, this new value is remembered as the new preset value for the next question. Higher bets will result in higher gains when the answer is correct, and lower points when the answer is wrong. Since Wordrobe is designed to create gold standard annotations, the correct choice is not defined (this is exactly what we want to obtain!). Therefore, the points are calculated on the basis of the answers given by other players, as in Phrase Detectives (Chamberlain et al., 2008). The idea is that the majority rules, meaning that the choice that gets selected most by human players is probably the correct one. So, the more players agree with each other, the more points they gain. As a consequence, the score of a player is continually updated – even when the player is not playing – in order to take into account the answers provided by other players answering the same questions.

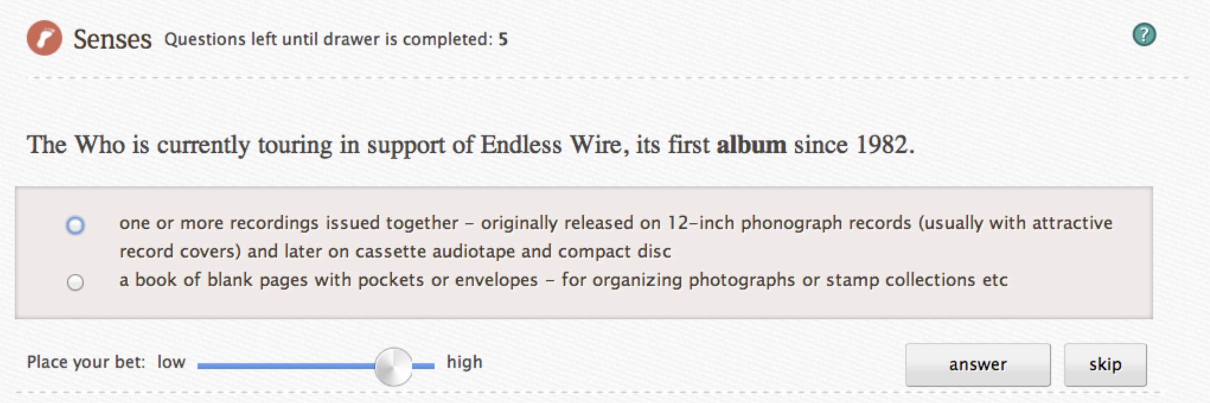
2.2 Generation of questions for the Senses game

All Wordrobe games consist of automatically generated multiple-choice questions. In the case of *Senses*, the word sense labeling game, each question consists of one or two sentences extracted from the Groningen Meaning Bank with one highlighted word for which the correct word sense in the given context must be determined. Currently, the game only focuses on nouns and verbs, but it can be easily extended to include, e.g., adjectives and adverbs. The choices for the questions are automatically generated from the word senses in WordNet (Fellbaum, 1998).

Of all the occurrences (tokens) of nouns and verbs in the GMB, 92.3% occurs in WordNet. This results in a total of 452,576 candidate questions for the *Senses* game. For the first version of Wordrobe, we selected a subset of the tokens that have at most five different senses in WordNet, such that the number of choices for each question is restricted. Figure 1 shows a screenshot of a question of *Senses*.

3 Results

The number of automatically generated questions for the first version of *Senses* was 3,121. After the first few weeks of Wordrobe going live, we had received 5,478 answers. Roughly half (1,673) of the questions received at least one answer, with an average of three answers per question. In order to test the validity of the method of using a GWAP to obtain reliable word sense annotations, we selected a subset of the questions with a reasonable response rate and created a gold standard annotation.



The Who is currently touring in support of Endless Wire, its first **album** since 1982.

○ one or more recordings issued together – originally released on 12-inch phonograph records (usually with attractive record covers) and later on cassette audiotape and compact disc

○ a book of blank pages with pockets or envelopes – for organizing photographs or stamp collections etc

Figure 1: Screenshot from Wordrobe game *Senses*.

3.1 Gold standard annotation

We created a gold standard annotation for a test set of 115 questions with exactly six answers each, which was used to evaluate the answers given by the players of Wordrobe. Four trained human annotators individually selected the correct sense for each of the target words in the test set. Fleiss's kappa was calculated to evaluate inter-annotator agreement, resulting in $\kappa = 0.79$, which is generally taken to reflect substantial agreement. Unanimity was obtained for 64% of the questions and 86% of the questions had an absolute majority vote. In a second step of evaluation, the non-unanimous answers were discussed between the annotators in order to obtain 100% agreement on all questions, the result of which was used as the gold standard annotation.

3.2 Agreement measures

Given a question and a set of player answers, we need a procedure to decide whether to accept a particular choice into our annotated corpus. One important factor is agreement: if a great majority of players agrees on the same choice, this choice is probably the correct one. Smaller majorities of players are more likely to be wrong. Another important factor is the number of answers: the more players have answered a question, the more we can presumably rely on the majority's judgement. In this work, we focus on the first factor (agreement) because the average answer rate per question is quite low throughout our data set. We tested a couple of simple agreement measures that determine whether a choice is counted as a winning answer. We measure recall and precision for each measure with respect to the gold standard.

The simplest measure accepts every choice that has a relative majority. It always accepts some choice, unless the two choices with the most answers are tied. A stricter measure ("absolute majority") accepts only the choices that were chosen by at least a certain fraction of players who answered the question, with some threshold $t \geq 0.5$. We used the values 0.5, 0.7 and 1.0 as threshold, the latter only accepting choices unanimously picked by players.

The measures described above simply choose the majority answer relative to some threshold, but fail to take into account the total number of players that answered the question and the number of possible choices for a question. These factors will become more important when we evaluate questions with a higher number of answers. We need a measure that determines whether the majority answer is chosen significantly more often than the other answers. This means that the answers should be significantly skewed towards one answer. In order to test such an effect, we can use Pearson's chi-square test, which determines the goodness-of-fit of a given distribution relative to a uniform distribution. If we take the distribution of answers over the set of possible choices, we can say that only those questions for which this distribution significantly differs from a uniform distribution ($p < 0.05$) are considered to provide an acceptable answer. Because the number of answers per question in our test set is relatively small, a significant result means that there is one choice towards which the answers accumulate. Determining which choice this is can accordingly be done using the relative-majority measure described above.

3.3 Evaluation

We evaluate the annotations obtained from Wordrobe by comparing the data of the test set (115 questions) to the gold standard. We used each of the agreement measures described above to select the answers with a high enough majority, and calculated precision (the number of correct answers with respect to the total number of selected answers), recall (the number of correct answers with respect to the total number of questions), and the corresponding F-score. The results are shown in Table 1.

Table 1: Precision and recall based on different agreement measures

Strategy	Precision	Recall	F-score
Relative majority	0.880	**0.834**	**0.857**
Absolute majority ($t = 0.5$)	0.882	0.782	0.829
Absolute majority ($t = 0.7$)	0.945	0.608	0.740
Unanimity ($t = 1$)	**0.975**	0.347	0.512
Chi-square test ($p < 0.05$)	0.923	0.521	0.666

As expected, the highest recall is obtained using the relative majority measure since this measure is the least conservative in accepting a majority choice. As the threshold for accepting a choice is set higher, recall drops and precision rises, up to a very high precision for the unanimity measure, but with a significant loss in recall. The measure based on Pearson's chi-square test is similar in being conservative; having only six answers per question in the test set, only the questions that are very skewed towards one choice give a significant result of the chi-square test.

As described above, each answer is associated with a bet between 10% and 100% of the points available for a question, which players can adjust based on how certain they are about their answer. The distribution of bets over all answers shows two significant peaks for these extremes: in 66% of the cases the maximum bet was chosen, and the default minimum bet was chosen in 12% of the cases. The main motivation for inserting the betting function was to be able to identify questions that were more difficult for players by looking for low bets. We tested the correlation between the average bet per question and the relative size of the majority (indicating agreement between players) over all questions using Pearson's product-moment correlation and found a small but significant positive effect ($r = 0.150$, $p < 0.01$). We expect that this effect will increase if more data is available.

In order to test whether questions with high average bets were easier, we repeated the evaluation, including only questions with a high average bet: $\bar{b} \geq 80\%$ (see Table 2). Recall is reduced strongly, as one would expect, but we do observe an increase in precision for all measures except unanimity. This higher precision suggests that indeed the results of the questions for which players on average place a high bet are more similar to the gold standard. However, we will need more data to confirm this point.

Table 2: Precision and recall based on different agreement measures for questions with $\bar{b} \geq 80\%$

Strategy	Precision	Recall	F-score
Relative majority	0.917	**0.478**	**0.629**
Absolute majority ($t = 0.5$)	0.930	0.461	0.616
Absolute majority ($t = 0.7$)	0.956	0.383	0.547
Unanimity ($t = 1$)	**0.961**	0.217	0.355
Chi-square test ($p < 0.05$)	0.950	0.330	0.355

4 Discussion

The goal of Wordrobe is to obtain annotations from non-expert annotators that are qualitatively close to gold standard annotations created by experts. This requires automatic techniques for filtering out low-quality answers. We evaluated the results obtained using some simple selection techniques with respect to a gold standard created by experts. We found that even with very conservative settings, optimizing for precision, we could still get a reasonably high recall (0.347). The highest precision, obtained using this most conservative measure (unanimity), was 0.975. In fact, a closer look at the data showed that there was exactly one question on which the choice unanimously picked by players differed from the gold standard annotation. This question is shown in (1).

(1) Although the last Russian **troops** left in 1994, the status of the Russian minority (some 30% of the population) remains of concern to Moscow.

a. soldiers collectively (synonyms: military personnel, soldiery)
b. a group of soldiers
c. a cavalry unit corresponding to an infantry company
d. a unit of Girl or Boy Scouts (synonyms: troop, scout troop, scout group)
e. an orderly crowd (synonyms: troop, flock)

While according to the gold standard annotation the correct answer was (1b), the six players who answered this question in the game unanimously chose (1a) as the correct answer. This example illustrates the difficulty of the task at hand very well; one could argue for the correctness of both of the possible answers. In this case, the average bet posed by the players (83%) is not helpful either in determining the difficulty of the question. This example suggests that using a more fine-grained gold standard annotation, with a ranking rather than selection of possible answers, may result in higher quality results.

Overall, the measures for calculating agreement show high numbers for precision, which were improved even more by only taking into account the questions that received a high average bet. The main drawback for this evaluation procedure is the restricted average number of answers per question. Although the recall for the unanimity measure remains at an acceptable level for the test set, this number is likely to decrease severely for questions with a higher number of answers. On the other hand, the measure based on the chi-square test is expected to become more reliable in the case of a larger dataset. In general, the evaluation measures discussed in section 3 are very basic and not robust against small datasets or unreliable annotators. With the recent uprise of crowdsourcing platforms such as Amazon's Mechanical Turk, there has been a revived interest in the task of obtaining reliable annotations from non-expert annotators. Various methods have been proposed to model annotated data such that it can be used as a gold standard (see, e.g., Carpenter, 2008; Snow et al., 2008; Beigman Klebanov and Beigman, 2009; Raykar et al., 2010). The goal of this paper was to provide a general idea of the quality of the data that can be obtained using games with a purpose. However, creation of a proper gold standard will require the collection of more data, and the use of more advanced techniques to obtain reliable annotations. As a first step towards this goal, we will make the data used in this paper available online,[2] such that interested readers can perform their own evaluation methods on this data.

5 Conclusions and future work

In this paper we described and evaluated the first results about the use of a 'Game with a Purpose' for annotating word senses. Although the amount of data obtained for each question is still relatively small (the largest amount of answers given to a reasonably sized amount of questions was 6), the results on precision and recall compared to the gold standard annotation are promising. We proposed several measures for determining the winning answer of a question, and compared them with respect to the precision and recall results. In this paper we focused on obtaining high precision scores, because the goal of the project is to obtain gold standard annotations which can be used to improve the Groningen Meaning Bank (Basile et al., 2012). Future work will focus on obtaining larger amounts of data and evaluating the annotations as part of an integration into the GMB. Moreover, this method for obtaining annotations will be applied and evaluated with respect to other linguistic phenomena, such as named entity tagging, noun-noun compound interpretation, and co-reference resolution.

[2] http://gmb.let.rug.nl/

Acknowledgements

We'd like to thank Bob Carpenter for comments and discussion of the Wordrobe data, and the pointers he provided to related work. Obviously we also thank all Wordrobe players (962 so far) that generated so many answers (41,541 so far) in such a short time.

References

Akkaya, C., A. Conrad, J. Wiebe, and R. Mihalcea (2010). Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 195–203. Association for Computational Linguistics.

Artignan, G., M. Hascoët, and M. Lafourcade (2009). Multiscale visual analysis of lexical networks. In *13th International Conference on Information Visualisation*, Barcelona, Spain, pp. 685–690.

Basile, V., J. Bos, K. Evang, and N. J. Venhuizen (2012). Developing a large semantically annotated corpus. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Beigman Klebanov, B. and E. Beigman (2009). From annotator agreement to noise models. *Computational Linguistics 35*(4), 495–503.

Carpenter, B. (2008). Multilevel bayesian models of categorical data annotation. Tech. report, Alias-i.

Chamberlain, J., M. Poesio, and U. Kruschwitz (2008). Addressing the Resource Bottleneck to Create Large-Scale Annotated Texts. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 375–380. College Publications.

Fellbaum, C. (Ed.) (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.

Kilgarriff, A. and J. Rosenzweig (2000). Framework and results for English SENSEVAL. *Computers and the Humanities 34*(1), 15–48.

Raykar, V., S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni, and L. Moy (2010). Learning from crowds. *The Journal of Machine Learning Research 11*, 1297–1322.

Rumshisky, A., N. Botchan, S. Kushkuley, and J. Pustejovsky (2012). Word sense inventories by non-experts. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Snow, R., B. O'Connor, D. Jurafsky, and A. Ng (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 254–263. Association for Computational Linguistics.

Fitting, Not Clashing!
A Distributional Semantic Model of Logical Metonymy

Alessandra Zarcone[1], Alessandro Lenci[2], Sebastian Padó[3], Jason Utt[1]
[1]Universität Stuttgart, [2]Universität Heidelberg, [3]Università di Pisa
[1]zarconaa,uttjn@ims.uni-stuttgart.de,
[2]alessandro.lenci@ling.unipi.it, [3]pado@cl.uni-heidelberg.de

Abstract

Logical metonymy interpretation (e.g. *begin the book* → *writing*) has received wide attention in linguistics. Experimental results have shown higher processing costs for metonymic conditions compared with non-metonymic ones (*read the book*). According to a widely held interpretation, it is the type clash between the event-selecting verb and the entity-denoting object (*begin the book*) that triggers coercion mechanisms and leads to additional processing effort. We propose an alternative explanation and argue that the extra processing effort is an effect of thematic fit. This is a more economical hypothesis that does not need to postulate a separate type clash mechanism: entity-denoting objects simply have a low fit as objects of event-selecting verbs. We test linguistic datasets from psycholinguistic experiments and find that a structured distributional model of thematic fit, which does not encode any explicit argument type information, is able to replicate all significant experimental findings. This result provides evidence for a graded account of coercion phenomena in which thematic fit accounts for both the trigger of the coercion and the retrieval of the covert event.

1 Introduction

Type clash in logical metonymy. Logical metonymy, also known as enriched composition (e.g. *The writer began the novel*), is generally explained in terms of a type clash between an event-selecting metonymic verb (*begin*) and an entity-denoting object (*novel*), triggering the recovery of a covert event (*writing*). Extensive psycholinguistic work (McElree, Traxler, Pickering, Seely, and Jackendoff (2001) and Traxler, Pickering, and McElree (2002), among others) has demonstrated extra processing costs for metonymic constructions. For example, Traxler et al. (2002) combine metonymic and non-metonymic verbs with entity-denoting and event-denoting nouns (*The boy [started/saw]$_V$ [the puzzle/fight]$_{NP}$*) and report significantly higher processing costs for the coercion combination (metonymic verb combined with entity-denoting object, e.g. *The boy started the puzzle*).

Building on this and similar experiments, Frisson and McElree (2008) ascribe the extra processing cost to “the deployment of operations to construct a semantic representation” of the event (*writing the novel*) that is supposed to be triggered by the type clash. However, this explanation remains problematic. Notably, metonymic interpretations are also possible for event-denoting objects given suitable contexts (e.g. *John is a wrestling fan, he really enjoyed the fight last night* → *watching the fight*).

Thematic fit in logical metonymy. Another pervasive aspect of language processing is thematic fit in the shape of selectional preferences, that is, expectations of predicative lemmas about plausible fillers for their argument slots (e.g., the fact that *eat* requires a *[+edible]* object or that *crook* is a more fitting object for *arrest* than *cop*). While early formalizations of selectional preferences aimed at modeling a binary distinction between “sensical” and “nonsensical” predicate-argument combinations, later work such as Wilks (1975) adopted a graded notion of selectional preference. In psycholinguistics, thematic fit has emerged as a pivotal concept to explain effects on expectations about upcoming input in language comprehension (McRae, Spivey-Knowlton, and Tanenhaus 1998; Ferretti, McRae, and Hatherell 2001; Matsuki, Chow, Hare, Elman, Scheepers, and McRae 2011).

Concerning logical metonymy, there is considerable behavioral as well as modeling evidence that thematic fit plays an important role in metonymy *interpretation*, that is, the retrieval of covert events for metonymical constructions. Behavioral studies (Zarcone and Padó 2011; Zarcone, Padó, and Lenci 2012) as well as computational models (Lapata and Lascarides 2003; Zarcone, Utt, and Padó 2012) found that the retrieved event will be the event most compatible with our knowledge about typical events and their participants (as captured, e.g. by generalized event knowledge, (McRae and Matsuki 2009)), that is the interpretation with the highest thematic fit with the context. This is in contrast to traditional accounts of logical metonymy (Pustejovsky 1995) which ascribe covert event retrieval to complex lexical entries associating entities with events corresponding to their typical function or creation mode (qualia: *book* → *read / write*). The advantage of thematic fit-based accounts is that they can account for the influence of context on interpretation. For example, given *baker* and *child* as subjects of *finish the icing*, *baker* will cue *spread* as a covert event, while *child* will cue *eat*, even though it is possible that bakers eat icing or that children spread it.

Thematic fit as a trigger of logical metonymy. In this paper, we propose that thematic fit can explain not only the interpretation phase of metonymy (that is implicit event recovery), but also to the *triggering phase*. We claim that thematic fit can provide a convincing explanation for the triggering of the coercion operation: metonymic verbs prefer event-denoting objects, and sentences that have traditionally been analysed involving a coercion operation have a low thematic fit between the verb and the object. This account preserves the observation that metonymic verbs disprefer entity-denoting objects, but can explain it purely in terms of standard graded accounts of selectional preferences[1]. A thematic-fit based account of logical metonymy would bring a clear advantage of theoretical economy: it accounts for two phenomena (triggering and interpretation) with a single mechanism, that is, generalized event knowledge (quantified in terms of thematic fit). Furthermore, generalized event knowledge / thematic fit operate in any type of predicate-argument composition (McRae and Matsuki 2009). Thus, an explanation in terms of thematic fit would bring metonymy closer to "normal" online language comprehension process.

We test this hypothesis by modeling three datasets from well-known experimental studies on metonymy. We compute thematic fit predictions of all items relying solely on distributional information, without any information about predicate semantic types, and compare differences in thematic fit across conditions with corresponding differences in processing cost from the experiments. As we show below, our distributional model of thematic fit predicts all significant effects that have been found in the experiments previous experiments and that were interpreted as type-clash effects.

2 Experimental Setup

2.1 A Distributional Model of Thematic Fit

Distributional semantic models (Turney and Pantel 2010) build on the Distributional Hypothesis (Harris 1954; Miller and Charles 1991) which states that the meaning of a word can be modelled by observing the contexts in which it is used. Current distributional models are typically built by collecting contexts of word occurrences in large corpora, where "context" can be defined in many possible ways. Pairwise word similarity is then computed by comparing the similarity between the vectors which record the word co-occurrences in the data. Distributional models have been successful in modelling a range of cognitive tasks, including lexical development (Li, Farkas, and MacWhinney 2004), category-related deficits (Vigliocco, Vinson, Lewis, and Garrett 2004), and thematic fit (Erk, Padó, and Padó 2010).

Distributional Memory (DM, Baroni and Lenci (2010)) is a general framework for building distributional semantic models from syntactically analysed corpora. It constructs a three-dimensional tensor of

[1] Similarly, thematic-fit based accounts of selectional preferences encompass binary distinctions (e.g., *eat* requires a *[+edible]* object), while still including more fine-grained differences (e.g., *crook* is a more fitting object for *arrest* than *cop*).

weighted word-relation-word tuples each tuple is mapped onto a score by a function $\sigma\colon \langle w_1\ r\ w_2\rangle \rightarrow \mathbb{N}$, where w_2 is a target word, r as a relation and w_1 an argument or adjuct of the target word. For example, ⟨*marine subj shoot*⟩ has a higher weight than ⟨*teacher subj shoot*⟩. The set of relations can be defined in different ways, which gives rise to different flavors of DM. We use TypeDM, which uses generic syntactic relations as well as lexicalized relations (see Baroni and Lenci (2010) for details).

For our experiments, we project the DM tensor onto a $W_1 \times RW_2$ matrix, and we represent each target word W_1 in terms of a vector with dimensions corresponding to pairs of context words and their relations ($R \times W_2$). On this matrix, we compute a verb's expectations for its most typical object with a method similar to Erk et al. (2010) and Lenci (2011): For each verb v, we determine the 20 highest-scoring nouns in object relation and compute the centroid $c_o(v)$ of their context vectors. The thematic fit of a new noun n for v's object position is then defined as the cosine of the angle between n's own context vector and $c_o(v)$. Since all the vectors' components are positive, the thematic fit values range between 0 and 1.

2.2 Datasets

As stated above, we model three datasets from psycholinguistic experiments. The datasets fall into two categories: sentence triplets, and sentence quadruplets. Please refer to the corresponding psycholinguistic studies (McElree et al. 2001; Traxler et al. 2002) for further details about how the datasets were built.

Sentence Triplets The two datasets in this group consist of sentence triplets: one sentence with a metonymic verb ("type-shift" in the original papers), paired with two non-metonymic sentences, where one condition shows high thematic fit and one low thematic fit ("preferred" and "non-preferred" in the original papers). An example is *the writer [finished / wrote / read]$_V$ the novel.*

McElree dataset. This dataset is composed of 31 triplets of sentences from the self-paced reading experiment in McElree et al. (2001), for a total of 93 sentences. We excluded two triplets from the original dataset for problems of coverage, as they included low-frequency words.

Traxler dataset 1. This dataset is composed of 35 triplets of sentences from the eye-tracking experiment (experiment 1) in Traxler et al. (2002), for a total of 105 sentences. We excluded one triplet from the original dataset for problems of coverage, as it included low-frequency words.

The finding for these materials was a main effect of verb type on reading times (McElree et al. 2001) and eye tracking times (Traxler et al. 2002). Pairwise comparisons in both studies yielded (a) higher processing costs for the metonymic condition and (b) no significant differences between the high- and low-typicality condition.

Sentence Quadruplets The dataset in this group consists of sentence quadruplets which cross two factors: (i) metonymic verb vs. non-metonymic verb and (ii) event-denoting object vs. entity-denoting object. As an example, consider *The boy [started / saw]$_V$ [the fight / puzzle]$_{NP}$.*

Traxler dataset 2. This dataset is composed of 32 sentence quadruplets from experiments 2 (eye-tracking) and 3 (self-paced reading) in Traxler et al. (2002), for a total of 120 sentences. We exclude one triplet from the original dataset for problems of coverage.

The findings for this dataset were a main effect of object type on eye tracking times (experiment 2) and on reading times (and experiment 3), and a significant verb$*$object interaction, with higher processing costs for the metonymic condition (metonymic verb combined with entity-denoting object).

2.3 Evaluation Method

For each test sentence, we compute the verb-object thematic fit in DM. We assume that processing load increases with lower thematic fit and that processing cost (reading time and eye tracking time) corresponds to $1 -$ thematic fit. We manipulate thematic fit in DM as a dependent variable, and we employ linear

		high-typicality	low-typicality	metonymic
triplets from McElree et al. (2001)	reading times at the obj. + 1 position	360	361	385
	$1 -$ thematic fit	0.484	0.571	0.763
triplets from Traxler et al. (2002)	eye tracking (total time)	397	405	444
	$1 -$ thematic fit	0.482	0.576	0.744

Table 1: Comparing behavioral data from McElree et al. (2001) and Traxler et al. (2002) and thematic fit data from the computational model

regression analyses to test for main effects of factors (object type, verb type) on the dependent variable, as well as Wilcoxon rank sum task to test the significance of pairwise differences between conditions in terms of thematic fit. We then verify if the same main effects and significant pairwise differences were yielded by the psycholinguistic models and by the computational model.

3 Results

Sentence Triplets On the McElree dataset, the model mirrors all effects reported by the experimental studies, namely by yielding a main effect of the object type ($F = 20.247, p < 0.001$), and significant differences between the metonymic condition and both high-typicality ($W = 877, p < 0.001$) and low-typicality ($W = 740, p < 0.001$) conditions, but no significant difference between the high- and low-typicality conditions ($W = 595, p > 0.5$).

The model also mirrors the experimental effects found for the Traxler dataset 1. It yields a main effect of the object type ($F = 18.084, p < 0.001$), and significant differences between the metonymic condition and both high-typicality ($W = 1050, p < 0.001$) and low-typicality ($W = 889, p < 0.01$) conditions, but only a marginal difference between the high- and low-typicality conditions ($W = 780.5, p = 0.5$).

Sentence Quadruplets On the Traxler dataset 2, the model mirrors the main effect of the object type ($F = 8.0039, p < 0.01$) and the verb$*$object type interaction ($F = 8.3455, p < 0.01$) reported both by the self-paced reading and by the eye tracking studies. This result is visualized in Figure 1, which shows the close correspondence between experimental results from self-paced reading and modeling results.

The model also yields the same pair-wise differences reported by the eye-tracking study: within the sentences with entity-denoting objects, metonymic verbs yield lower thematic fit compared to non-metonymic sentences ($W = 300, p < 0.05$). Within the sentences with metonymic verbs, entity-denoting objects also yield lower thematic fit compared to entity-denoting objects ($W = 208, p < 0.01$).

4 Discussion and Conclusions

The distributional models successfully replicate the pattern of results of the psycholinguistic experiments. For the triplet dataset, the model yields a main effect of verb type, and produces the lowest thematic fit for the metonymic condition (corresponding to the longest reading times and eye fixations in the experiments); also, the significant differences detected by the model are the same as those in the experimental studies: high-typicality vs. metonymic and low-typicality vs. metonymic. Interestingly enough, while the computational model does not reveal a significant difference between what we called high- and low-typicality conditions (preferred and non-preferred conditions according to the paper's terminology), the psycholinguistic experiments were not able to show one either, suggesting that in the experimental materials that were employed, the preferred and non-preferred conditions do not differ significantly with regard to typicality.

For the quadruplet dataset, the model yields a main effect of object type and a significant verb$*$object interaction, producing the lowest thematic fit for metonymic verbs combined with entity-denoting objects.

		metonymic verb		non-metonymic verb	
		EN obj.	EV obj.	EN obj.	EV obj.
quadruplets from Traxler et al. (2002)	self-paced reading times at obj. + 1	512	427	467	455
	1 − thematic fit	0.230	0.336	0.283	0.287

Table 2: Comparing behavioral data (self-paced reading) from Traxler et al. (2002) and thematic fit data from the computational model. Eye-tracking data mirror the findings of the self-paced reading study.

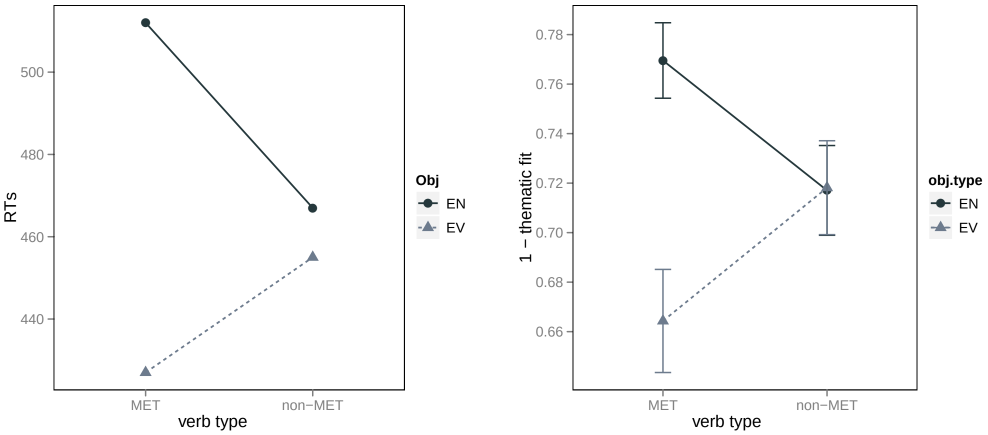


Figure 1: Comparing reading times in Traxler et al. (2002) (left, experiment 3) and results from thematic fit from our model (right).

The significant differences detected by the model were the same as those in the experimental studies: for the entity-denoting objects, metonymic verbs yielded lower thematic fit than non-metonymic verbs, whereas this difference was not significant for event-denoting objects; for the metonymic verbs, entity-denoting objects yielded lower thematic fit than event-denoting objects, whereas this difference was not significant for non-metonymic verbs.

The model's success in replicating the results from the psycholinguistic experiments shows that distributional similarity-based estimates of thematic fit are sufficient to account for the behavioral correlates of logical metonymy processing (such as longer reading and eye-tracking times) found so far in cognitive experiments. More precisely, the model, which did not encode any explicit type information, was able to replicate all significant experimental findings which were formulated based on type distinctions. This result supports the following general claims that we are currently testing on further experiments: i.) the two phenomena of triggering and interpretation in logical metonymy can be explained with a single mechanism relying on general event knowledge activation and integration; ii.) metonymy is actually much more closer to “normal” online predicate-argument composition processes, both being based on thematic fit computation. The structured distributional model of the latter proves to be an interesting tool to critically reanalyze psycholinguistic datasets, by (a) highlighting possible implicit lexical-semantic biases which may influence the results and (b) providing alternative explanations for them.

Acknowledgements The research for this paper was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the SFB 732 "Incremental specification in context" / project D6 "Lexical-semantic factors in event interpretation" at the University of Stuttgart.

References

Baroni, M. and A. Lenci (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics 36*(4), 1–49.

Erk, K., S. Padó, and U. Padó (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics 36*(4), 723–763.

Ferretti, T. R., K. McRae, and A. Hatherell (2001). Integrating verbs, situation schemas and thematic role concept. *Journal of Memory and Language 44*, 516–547.

Frisson, S. and B. McElree (2008). Complement coercion is not modulated by competition: Evidence from eye movements. *Journal of Exp. Psychology: Learning, Memory, and Cognition 34*(1), 1.

Harris, Z. S. (1954). Distributional structure. *Word 10*(23), 146–162.

Lapata, M. and A. Lascarides (2003). A probabilistic account of logical metonymy. *Computational Linguistics 29*(2), 263–317.

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of CMCL*, Portland, Oregon, pp. 58–66.

Li, P., I. Farkas, and B. MacWhinney (2004). Early lexical development in a self-organizing neural network. *Neural Networks 17*, 1345–1362.

Matsuki, K., T. Chow, M. Hare, J. L. Elman, C. Scheepers, and K. McRae (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Exp. Psychology: Language, Memory, and Cognition 37*(4), 913–934.

McElree, B., M. Traxler, M. Pickering, R. Seely, and R. Jackendoff (2001). Reading time evidence for enriched composition. *Cognition 78*(1), B17–B25.

McRae, K. and K. Matsuki (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass 3/6*, 1417–1429.

McRae, K., M. Spivey-Knowlton, and M. Tanenhaus (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language 38*(3), 283–312.

Miller, G. A. and W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*(1), 1–28.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Traxler, M. J., M. J. Pickering, and B. McElree (2002). Coercion in sentence processing: evidence from eye-movements and self-paced reading. *Journal of Memory and Language 47*, 530–547.

Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research 37*, 141–188.

Vigliocco, G., D. P. Vinson, W. Lewis, and M. F. Garrett (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology 48*(4), 422–488.

Wilks, Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence 6*(1), 53–74.

Zarcone, A. and S. Padó (2011). Generalized event knowledge in logical metonymy resolution. In *Proceedings of CogSci*, Austin, TX, pp. 944–949.

Zarcone, A., S. Padó, and A. Lenci (2012). Inferring covert events in logical metonymies: a probe recognition experiment. In *Proceedings of CogSci*, Austin, TX.

Zarcone, A., J. Utt, and S. Padó (2012). Modeling covert event retrieval in logical metonymy: Probabilistic and distributional accounts. In *Proceedings of CMCL*, Montreal, Canada.