Representing semantics of texts - a non-statistical approach

Svetlana Hensman
School of Computing
Dublin Institute of Technology
Kevin Street
Dublin 8, Ireland
Svetlana.Hensman@comp.dit.ie

John Dunnion
Intelligent Information Retrieval Group
Department of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
John.Dunnion@ucd.ie

Abstract

This paper describes a non-statistical approach for semantic annotation of documents by analysing their syntax and by using semantic/syntactic behaviour patterns described in VerbNet. We use a two-stage approach, firstly identifying the semantic roles in a sentence, and then using these roles to represent some of the relations between the concepts in the sentence and a list of noun behaviour patterns to resolve some of the unknown (generic) relations between concepts. All outlined algorithms were tested on two corpora which differs in size, type, style and genre, and the performance does not vary significantly.

1 Introduction

This paper describes a system for semi-automatic conceptual graph acquisition using a combination of linguistic resources, such as VerbNet and WordNet, together with semi-automatically compiled domain-specific knowledge. Such semantic information has a number of possible applications, for example in the area of information retrieval/extraction for enhancing the search methods or in question-answering systems, allowing users to communicate with the system in natural language (English).

We use conceptual graphs (CGs) (Sowa, 1984), a knowledge-representation formalism based on semantic networks and the existential graphs of C. S. Pierce, to represent the semantics of documents. There are number of systems for generating conceptual graphs representation of sentences: Sowa and Way (Sowa and Way, 1986) use a lexicon of canonical graphs which are combined to build a conceptual graph representation of a sentence, while Veraldi at al. (Velardi et al., 1988) describe a prototype of a semantic processor for Italian sentences, which uses a manually acquired lexicon of about 850 word-sense definitions, each including 10 – 20 surface semantic patterns (SSPs) representing both usage information and semantic constraints.

There are also systems aimed at extracting partial knowledge from texts, by either filling semantic templates (Hobbs et al., 1996) or by generating a set of linguistic patterns for information extraction (Harabagiu and Maiorano, 2000), to name but a few.

The following sections describe in more detail the various aspects of our system, the experiments that we carried out to test the proposed algorithms and finally draw some conclusions.

2 System overview

We use a two-step approach for constructing conceptual graph representations of texts: firstly, by using VerbNet and WordNet, we identify the semantic roles in a sentence, and secondly, using these semantic roles and a set of syntactic/semantic rules we construct a conceptual graph.

To evaluate our algorithms we use test documents from two corpora in different domains – the Reuters-21578 text categorization test collection (Reuters, 1987) and the collection of aviation incident reports provided by the Irish Air Accident Investigation Unit (AAIU) (Air Accident Investigation Unit, 2004). All documents are parsed using Eugene Charniak's maximum entropy inspired parser (Charniak, 2000).

3 Semantic role identification

There are number of different existing approaches for identifying semantic roles, varying from traditional parsing approaches, for example using HPSG grammars and Lexical Functional Grammars, that strongly rely on manually-developed grammars and lexicons, to data-driven approaches, for example AutoSlog (Riloff and Schmelzenbach, 1998). In the domain of the Air Traveler Information System (Miller et al., 1996) the authors apply statistical methods to compute the probability that a constituent can fill in a semantic slot within a semantic frame. Gildea and Jurafsky (Gildea and Jurafsky, 2002) describe a statistical approach for semantic role labelling using data collected from FrameNet by analysing a number of features such as *phrase type*, *grammatical function*, *position* in the sentence, etc.

Shi and Mihalcea (Shi and Mihalcea, 2004) propose a rule-based approach for semantic parsing using FrameNet and WordNet. They extract rules from the tagged data provided by FrameNet, which specify the realisation (order and different syntactic features) for the present semantic roles. They also create a feature set representation of the sentence and match it to each of the extracted rules. The result is the rule providing the most feature matches. The authors do not provide any information on how they select between different matches with the same score, or if there is any semantic check on suitability of a phrase to realise a semantic role (FrameNet does not provide any restrictions on the semantic roles similar to the selectional restrictions present in VerbNet).

The approach we propose for semantic role identification uses information about each verb's behaviour, provided in VerbNet, and the WordNet taxonomy to decide whether a phrase can be a suitable match for a semantic role.

VerbNet (Kipper et al., 2000) is a computational verb lexicon, based on Levin's verb classes, that contains syntactic and semantic information for English verbs. Each VerbNet class defines a list of *members*, a list of possible *thematic roles*, and a list of *frames (patterns)* of how these semantic roles can be realized in a sentence.

WordNet (Fellbaum, 1998) is an English lexical database containing about 120 000 entries of nouns, verbs, adjectives and adverbs, hierarchically organized in synonym groups (called *synsets*), and linked with relations such as *hypernym*, *hyponym*, *holonym* and others.

To identify the semantic roles for a clause in a sentence we identify and match the clause pattern to each of the possible semantic frames for the clause verb (from VerbNet). The result is a list of all possible semantic role assignments, from which we must identify the correct one.

3.1 Constructing sentence patterns for the verbs in a sentence

For each sentence clause we construct a syntactical pattern, which is a flat parse representation that identifies the main verb and the other main categories of the clause. As a sentence can have subordinate clauses, we usually have more than one syntactic pattern per sentence. Each such pattern is processed individually.

Using a constituency parser (such as Charniak's) is suitable in the majority of cases, but there are some sentences where the correct set of role fillers cannot be identified by using the parse tree. For example, for sentences such as

> *The price of oil will rise by 5 cents by the end of the year.*

the phrase *the price of oil* will be identified as a possible role filler by our system, while the correct result would have *the price* identified as the *Attribute* and *oil* as the *Patient*. For such cases the use of a dependency parser (such as a Link Grammar parser or a Functional Dependency Grammar parser) would be required.

We also address some simple cases of pronoun anaphoric reference. For example, for patterns such as

> *Iomega Corp said it has laid off over a quarter of its professional and management staff.*

we identify the pronoun *it* as referring to the subject of the verb in the main clause (which here is *Iomega Corp*) if they agree by gender and number. In cases where the type of the concept represented by the phrase is known, an agreement by type is also required.

Some cases of intersentential pronoun anaphoric references are also resolved by analysing the previous sentence context for suitable candidates, that agree by gender, number and type. Agreement by type is present if the type of the phrase is compatible (or the same) as the type of the phrase it references. For example, if *the company* refers to *Iomega Corp*, which is listed as an instance of the type *organization*, then the types of the two phrases are compatible, as *company* is defined as sub-type of *organization*. If agreement by type cannot be assured, the reference is not resolved. The reference is only resolved if there is a single possibility for its resolution.

3.2 Extracting VerbNet semantic role frames

Each verb can be described in VerbNet as a member of more than one class, and therefore the list of its possible semantic frames is a combination of the semantic frames defined in each of the classes in which it participates.

We extract all the semantic frames in a class and consider them to be possible semantic frames for each of the verbs that are members of this class. Each verb class also defines a list of selectional constraints for the semantic roles. For example, for all the verbs that are members of the VerbNet class **get-13.5.1** one of the possible semantic role frames is:

Agent[+animate OR +organization] V Theme Prep(from) Source[+concrete].

The selectional constraints check is implemented using one or a combination of the following techniques: hypernym relations defined in WordNet, pattern matching techniques, syntactic rules and some heuristics.

3.3 Matching algorithm

The matching algorithm matches the sentence pattern against each of the possible semantic role frames extracted from VerbNet. We match the constituents before and after the verb in the sentence pattern to the semantic roles before and after the verb in the semantic role frame.

If the number of the available constituents in the sentence pattern is less than the number of the required slots in the frame, the match fails. If there is more than one constituent available to fill a slot in a semantic frame, each of them is considered a different match. If, for a semantic frame, we find a constituent for each of the semantic role slots that complies with the selectional constraints, the algorithm considers this a possible match.

Multiple results are identified when there are two or more phrases in a sentence that are possible semantic role realisations, or if there are two or more semantic frames for which matches were found. To select the correct role assignment we use a weighting function that assigns scores to each result and returns the one with the highest score. For each identified role the weighting function adds one point if the role does not have any selectional restrictions, and two points if there are restrictions (including prepositional restrictions). The total score for a solution is the sum of the scores for each identified roles. The solution with the highest score is selected.

For example, for the sentence

> *USAir bought Piedmont for 69 dlrs cash per share.*

the algorithm identifies two possible role assignments:

> *Agent[+animate OR +organization]* matching NP(The company)
> *Theme[]* matching NP(the shares)
> *Asset[+currency]* matching PP(for 69 dlrs cash per share)

with $weight_{frame1}$ = *2 + 1 + 2 = 5* and the second solution

> *Agent[+animate OR +organization]* matching NP(The company)
> *Theme[]* matching the NP(the shares)

with $weight_{frame5}$ = *2 + 1 = 3*

Therefore, the algorithm returns the first set of role assignments as a result.

4 Building conceptual graphs

The conceptual graph representation of the sentence is built through the following steps: firstly, for each of the constituents of the sentence we recursively build a conceptual graph representation; then we link all the conceptual graphs representing the constituents into a single graph; and finally, we resolve the unknown (generic) relations. Each of these steps is described in more detail in the following sub-sections.

4.1 Building a conceptual graph representation of a phrase

The first step involves building a conceptual graph for a phrase. Our general assumption is that each lexeme in the sentence is represented using a separate concept, therefore all nouns, adjectives, adverbs and pronouns are represented using concepts, while the determiners and numbers are used to specify the referent of the relevant concept (thus further specifying the concept).

Below we illustrate the procedure for building a conceptual graph for some of the most common types of phrases.

- **NP -> DT JJ NN**

  For phrases following this pattern we create two concepts - one for the NN with a concept referent corresponding to the type of the determiner DT, and another concept representing the adjective, and link both of them by an *Attribute* relation. If the phrase contains more than one adjective, each of them is represented by a separate concept and they are all linked with *Attribute* relations to the concept representing the noun.

- **NP -> NP , SBAR ,**

  This pattern represents phrases where the noun is further specified by the SBAR (for example, ***The co-pilot, who was acting as a main pilot,*** *landed the plane.*) For these patterns a conceptual graph is built for the SBAR and the head concept, if a WHNP phrase (e.g. *which* or *who*), is replaced by the concept created for the NP.

- **PP -> IN NP**

  For such prepositional phrases we construct a conceptual graph representing the noun phrase. We also keep track of the preposition heading the prepositional phrase, as it is used to mark the relation between this phrase and the other relevant phrases in the sentence.

4.2 Attaching all constituents to the verb

Once the graphs for each of the constituents are constructed they are linked together in a single conceptual graph. As each of them describes some aspect of the concept represented by the verb, we link them to that concept.

If the constituent already has an identified semantic role during the previous phase, the same relation is used when constructing the conceptual graph between the CG representing the constituent and the verb. If the constituent does not have any semantic roles identified, a relation with a generic label is used, which allows us to build the structure of the CG concentrating on the concepts involved, and to resolve the generic labels at a later stage. The generic labels used are either *REL*, or in the case of prepositional phrases headed with a proposition *prep*, *REL_prep* (e.g. *REL_on*).

4.3 Resolving unknown relations

Finally we resolve some of the unknown (generic) relations in the conceptual graph. We keep a database of the most common syntactic realisation of relations between concepts with specific types. An example of a relation correction rule is:

Flight ***REL_from*** *City* -> *Flight* **Source** *City*

where the left part of the rule represents the two concepts linked by a generic relation and the right side represents the graph after the modification.

All generic relations present after this step must be manually resolved by the user. The system offers help by suggesting possible relations introduced by a preposition. For example, the preposition *for* can indicate *Beneficiary* (e.g. *a book* ***for Mary***), *Duration* (e.g. ***for three hours***), etc.

5 Query representation

Representation of questions differs than representation of declarative sentences and deserves special attention. For sentences representing questions we try to identify the statement that will correspond to the question and then construct the conceptual graph in a similar way as for declarative sentences.

5.1 Yes/No questions

We process simple yes/no questions (questions that require a yes/no answer) that are constructed by a subject-verb inversion by applying a transformation to reverse the question to a declarative sentence.

5.2 Wh_question

The parse tree of a sentence expressing a wh_question has the following general structure: *SBARQ ->WH_phrase SQ ?* where the *WH_phrase* is either *WHNP*, *WHADVP* or *WHPP* and represents the concept that triggers the query. The *SQ* represents the rest of the sentence.

Similarly to yes/no questions, these type of questions are also transformed to declarative sentences. The *wh_word* (e.g. *who*, *what*, *where*, *when*) is represented by a generic concept. The relation that attaches this concept to the verb depends on the type of the wh_phrase and can be one of the following:

WHNP

These phrases are headed by the wh_ question words *who*, *what* or *which*. The relation between the wh_phrase and the verb is either identified from applying a suitable semantic frame for this verb, or it is a generic one, *REL*.

WHADVP

These phrases represent an adverbial modifier for time, place or location. If the phrase marked as *WHADVP* is *where* the relation is *locative*; if it is *when*, the relation is *temporal*; and if it is *how*, the relation is *manner*.

WHPP

Such phrases are not processed by our system.

6 Experimental results

Each module of the system was evaluated separately.

The first experiment we carried out was to estimate the accuracy of the sentence frame constructed by the role labelling module and it was performed on randomly selected 2% of the verbs in Reuters and 7% of the verbs in AAIU corpora. The parse trees produced by Charniak's parser were manually edited to avoid any errors due to incorrect parses. The results showed that the system identified the correct set of possible candidates for semantic roles for 90% and 89% of the verbs in the Reuters and in the AAIU documents respectively.

Further experiments were carried out to evaluate the performance of the role assigning module. As a testbed we randomly selected 2% of the verbs in Reuters and 15% of the verbs in the AAIU documents. From these, we analysed only those cases where the verb is a member of at least one VerbNet frame and the possible role candidates were correctly identified. For 60% and 70% of the remaining verbs, respectively, the algorithm identifies a single correct solution. In 3% and 4% of the cases respectively a partially correct result is found (in the majority of such cases it is Agent, Patient and Theme roles that are correctly identified, together with some incorrect ones).

In 11% and 9% of the cases for Reuters and AAIU, respectively, the algorithm identifies a set of possible solutions, containing the correct and several incorrect ones. For these cases the weighting function identifies the correct solution in 38% of the of the cases for AAIU documents and 59% of the cases for the Reuters documents, while in 40% and 21% of the cases, respectively, it identifies the correct and one or more incorrect results.

We also evaluated the percentage of the syntactic patterns that the graph builder recognises: for AAUI and Reuters documents, respectively, we can build a graph for 76% and 67% of the noun phrases, for 95% and 94% of the prepositional phrases and for 91% and 97% of the subordinate clauses.

7 Conclusions

In this paper we have described an approach for constructing conceptual graphs for English sentences, using VerbNet, WordNet and some domain-specific knowledge. The achieved accuracy is strongly influenced by the lack of VerbNet descriptions of many verbs present in both corpora, as well as the lack of semantic frames for the present verb sense. Also, as the approach is not statistical, it does not require large amount of training data.

There are several other lexical resources currently available that seem suitable for semantic role identification, among them FrameNet and PropBank. Our choice of VerbNet as a lexical resource is based on our belief that a set of domain-independent descriptive role labels (such as those defined in VerbNet) is more suitable as it allows for generalisations.

A drawback of both FrameNet and PropBank is that the roles do not include any selectional restrictions, which makes it hard for a non-statistical method to identify the correct filler for each role. As shown earlier, the selectional restrictions defined for the semantic roles prove to be a valuable asset when deciding if a phrase can be a role filler. While we can resolve the majority of them by analysing the syntactic structure or by using the WordNet hierarchy, some are more difficult to resolve. For example, the restriction *solid* describes an attribute or a state of an object, relations which cannot be checked by using WordNet.

FrameNet on the other hand defines usages not only for verbs, but also for nouns. As one of the causes for the relatively poor performance of the conceptual graph building module is the lack of a sufficient number of relation-correction rules, our current approach to increasing their number is trying to extract the rules from FrameNet.

Work on the system is ongoing and efforts are continuing to implement a verb sense disambiguation component.

References

Air Accident Investigation Unit. 2004. Irish Air Accident Investigation Unit Reports. Available online: (http://www.aaiu.ie/).

Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL-2000*, pages 132–139.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, May.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

Sanda Harabagiu and Steven Maiorano. 2000. Acquisition of linguistic patterns for knowledge-based information extraction. In *Proceedings of LREC-2000*, Athens, June.

Jerry Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson. 1996. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In *In Finite State Devices for Natural Language Processing*, Cambridge, MA. MIT Press.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *Proceedings of Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 691 – 696, Austin, TX, July 30 - August 3.

Scott Miller, David Stallard, Robert Bobrow, and Richard Schwartz. 1996. A fully statistical approach to natural language interfaces. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 55–61, Santa Cruz, June. Morgan Kaufmann Publishers, Inc.

Reuters. 1987. Reuters-21578 Text Categorization Collection. Available online: (http://kdd.ics.uci.edu/databases/reuters21578/-reuters21578.html).

Ellen Riloff and Mark Schmelzenbach. 1998. An Empirical Approach to Conceptual Case Frame Acquisition. In *Proceedings of the Sixth Workshop on Very Large Corpora*.

Lei Shi and Rada Mihalcea. 2004. Open Text Parsing Using FrameNet and WordNet. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *Proceedings of HLT-NAACL 2004: Demonstration Papers*, pages 247–250, Boston, Massachusetts, USA, May 2 – May 7. Association for Computational Linguistics.

John F. Sowa and Eileen C. Way. 1986. Implementing a semantic interpreter using conceptual graphs. *IBM Journal of Research and Development*, 30(1):57–69, January.

John F. Sowa. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading, M.

Paola Velardi, Maria Teresa Pazienza, and Mario DeGiovanetti. 1988. Conceptual graphs for the analysis and generation of sentences. *IBM Journal of Research and Development*, 32(2):251–267, March.