Automatic Enrichment of WordNet with Common-Sense Knowledge

Luigi Di Caro, Guido Boella
University of Turin
Corso Svizzera 185 - Turin - Italy
dicaro@di.unito.it, boella@di.unito.it

Abstract

WordNet represents a cornerstone in the Computational Linguistics field, linking words to meanings (or senses) through a taxonomical representation of synsets, i.e., clusters of words with an equivalent meaning in a specific context often described by few definitions (or glosses) and examples. Most of the approaches to the Word Sense Disambiguation task fully rely on these short texts as a source of contextual information to match with the input text to disambiguate. This paper presents the first attempt to enrich synsets data with common-sense definitions, automatically retrieved from ConceptNet 5, and disambiguated accordingly to WordNet. The aim was to exploit the *shared- and immediate-thinking* nature of common-sense knowledge to extend the short but incredibly useful contextual information of the synsets. A manual evaluation on a subset of the entire result (which counts a total of almost 600K synset enrichments) shows a very high precision with an estimated good recall.

Keywords: Semantic Resources, Semantic Enrichment, Common-Sense Knowledge

1. Introduction

In the last 20 years, the Artificial Intelligence (AI) community working on Computational Linguistics (CL) has been using one knowledge base among all, i.e., WordNet (Miller, 1995). In few words, WordNet was a first answer to the most important question in this area, which is the treatment of language ambiguity.

Generally speaking, a word is a symbolic expression that may refer to multiple meanings (polysemy), while distinct words may share the same meaning. Syntax reflects grammar rules which add complexity to the overall communication medium, making CL one of the most challenging research area in the AI field.

From a more detailed perspective, WordNet organizes words in *synsets*, i.e., sets of words sharing a unique meaning in specific contexts (synonyms), further described by descriptions (*glosses*) and examples. Synsets are then structured in a taxonomy which incorporates the semantics of generality/specificity of the referenced concepts. Although extensively adopted, the limits of this resource are sometimes critical: 1) the top-down and general-purpose nature at the basis of its construction lets asking about the actual need of some underused meanings, and 2) most Word Sense Disambiguation approaches use WordNet glosses to understand the link between an input word (and its context) and the candidate synsets.

In recent years, natural language understanding traced the line of novel and challenging research directions which have been unveiled under the name of textual entailment and question answering. The former is a form of inference based on a lexical basis (to be not intended as in Formal Semantics) while the latter considers the last-mile goal of every computational linguists' dream: asking questions to a computer in natural language as with humans.

As a matter of fact, these tasks require an incredibly rich semantic knowledge containing facts related to behavioural rather than conceptual information, such as what an object may or may not do or what may happen with it after such actions.

In the light of this, an interesting source of additional gloss-like information is represented by common-sense knowledge (CSK), that may described as a set of shared and possibly general facts or views of the world. Being crowd-sourced, CSK represents a promising (although often complex and uncoherent) type of information which can serve complex tasks such as the ones mentioned above. ConceptNet is one of the largest sources of CSK, collecting and integrating data from many years since the beginning of the MIT Open Mind Common Sense project. However, terms in ConceptNet are not disambiguated, so it is difficult to use due to its large amount of lexical ambiguities.

To make a parallelism, this double-edged situation is similar to a well-known research question in the Knowledge Representation and Information Retrieval fields, regarding the dichotomy between taxonomy and folksonomy. The former is a top-down and often human-generated representation of a domain whereas the latter comes from free tags associated to objects in different contexts. A large effort in the integration of these two types of knowledge has been carried out, e.g., in (Collins and Murphy, 2013) (Di Caro et al., 2011) (Kiu and Tsui, 2011).

This paper presents a novel method for the automatic enrichment of WordNet with disambiguated semantics of ConceptNet 5. In particular, the proposed technique is able to disambiguate common-sense instances by linking them to WordNet synsets. A manual validation of 505 random enrichments shown promising a 88.19% of accuracy, demonstrating the validity of the approach.

2. Related Work

The idea of extending WordNet with further semantic information is not new. Plenty of methods have been proposed, based on corpus-based enrichments rather than by means of alignments with other resources.

In this section we briefly mention some of the most relevant but distinct works related to this task: (Agirre et al., 2000) use WWW contents to enrich synsets with topics (i.e., sets of correlated terms); (Ruiz-Casado et al., 2007) and (Navigli and Ponzetto, 2010) enrich WordNet with Wikipedia and other resources; (Montoyo et al., 2001) use a Machine

Learning classifier trained on general categories; (Navigli et al., 2004) use OntoLearn, a Word Sense Disambiguation (WSD) tool that discovers patterns between words in WordNet glosses to extract relations; (Niles and Pease, 2003) integrates WordNet with the Sumo ontology (Pease et al., 2002); (Bentivogli and Pianta, 2003) extends WordNet with phrases; (Laparra et al., 2010) align synsets with FrameNet semantic information; (Hsu et al., 2008) combine WordNet and ConceptNet knowledge for expanding web queries; (Chen and Liu, 2011) align ConceptNet entries to synsets for WSD; (Bentivogli et al., 2004) integrates WordNet with domain-specific knowledge; (Vannella et al., 2014) extends WordNet via games with a purpose; (Niemi et al., 2012) proposed a bilingual resource to add synonyms.

3. Common-sense Knowledge

The Open Mind Common Sense[1] project developed by MIT collected unstructured common-sense knowledge by asking people to contribute over the Web. In this paper, we started focusing on ConceptNet (Speer and Havasi, 2012), that is a semantic graph that has been directly created from it. In contrast with the mentioned linguistic resources, ConceptNet contains common-sense facts which are conceptual properties but also behavioural information, so it perfectly fits with the proposed model. Examples of properties are *partOf, madeOf, hasA, definedAs, hasProperty* and examples of functionalities are *capableOf, usedFor, hasPrerequisite, motivatedByGoal, desires, causes*.

Among the more unusual types of relationships (24 in total), it contains information like "*ObstructedBy*" (i.e., referring to what would prevent it from happening), "and *CausesDesire*" (i.e., what does it make you want to do). In addition, it also has classic relationships like "*is_a*" and "*part_of*" as in most linguistic resources (see Table 1 for examples of property-based and function-based semantic relations in ConceptNet). For this reason, ConceptNet has a wider spectrum of semantic relationships but a much more sparse coverage. However, it contains a large set of function-based information (e.g., all the actions a concept can be associated with), so it represents a good basis for a complementary enrichment of WordNet.

4. The Proposed Approach

This paper proposes a novel approach for the alignment of linguistic and common-sense semantics based on the exploitation of their intrinsic characteristics: while the former represents a reliable (but strict in terms of semantic scope) knowledge, the latter contains an incredible wide but ambiguous set of semantic information. In the light of this, we assigned the role of *hinge* to WordNet, that guides a trusty, multiple and simultaneous retrieval of data from ConceptNet which are then intersected with themselves through a set of heuristics to produce automatically-disambiguated knowledge. ConceptNet (Speer and Havasi, 2012) is a semantic graph that has been directly created from the Open Mind Common Sense[2] project developed by MIT, which collected unstructured common-sense knowledge by asking people to contribute over the Web.

[1] http://commons.media.mit.edu/

[2] http://commons.media.mit.edu/

4.1. Common-sense Data: ConceptNet 5

The Open Mind Common Sense[3] project developed by MIT collected unstructured common-sense knowledge by asking people to contribute over the Web. In this paper, we make use of ConceptNet (Speer and Havasi, 2012), that is a semantic graph that has been directly created from it. In contrast with linguistic resources such the above-mentioned WordNet, ConceptNet contains semantics which is more related to common-sense facts.

4.2. Basic Idea

The idea of the proposed enrichment approach relies on a fundamental principle, which makes it novel and more robust w.r.t the state of the art. Indeed, our extension is not based on a similarity computation between words for the estimation of correct alignments. On the contrary, it aimed at enriching WordNet with semantics containing direct relations- and words overlapping, preventing associations of semantic knowledge on the unique basis of similarity scores (which may be also dependent on algorithms, similarity measures, and training corpora). This point makes this proposal completely different from what proposed by (Chen and Liu, 2011), where the authors created word sense profiles to compare with ConceptNet terms using semantic similarity metrics.

4.3. Definitions

Let us consider a WordNet synset $S_i =< T_i, g_i, E_i >$ where T_i is the set of synonym terms $t_1, t_2, ..., t_k$ while g_i and E_i represent its gloss and the available examples respectively. Each synset represents a meaning ascribed to the terms in T_i in a specific context (described by g_i and E_i). Then, for each synset S_i we can consider a set of semantic properties $P_{wordnet}(S_i)$ coming from the structure around S_i in WordNet. For example, *hypernym*(S_i) represents the direct hypernym synset while *meronyms*(S_i) is the set of synsets which compose (as a *made-of* relation) the concept represented by S_i. The above-mentioned complete set of semantic properties $P_{wordnet}(S_i)$ of a synset S_i contains a set of pairs $< rel-word >$ where rel is the relation of S_i with the other synsets (e.g., *is-a*) and $word$ is one of the lemmas of such linked synsets. For example, given the synset S_{cat} : *cat, true cat (feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats*), one resulting $< rel - word >$ pair that comes from *hypernym*(S_{cat}) will be $< isA - feline >$ since $feline$ is one lemma of the hypernym synset $S_{feline,felid}$: *feline, felid (any of various lithe-bodied roundheaded fissiped mammals, many with retractile claws)*. Note that in case of multiple synonym words in the related synsets, there will be multiple $< rel - word >$ pairs. Then, ConceptNet can be seen as a large set of semantic triples in the form $NP_i - rel_k - NP_j$, where NP_i and NP_j are simply non-disambiguated noun phrases whereas rel_k is one of the semantic relationships in ConceptNet.

4.4. Algorithm and heuristics

At this point, the problem is the alignment of ConceptNet triples with WordNet synsets. For this reason, the algorithm

[3] http://commons.media.mit.edu/

Table 1: Some of the existing properties (P) and functionalities (F) in ConceptNet, with example sentences in English.

Relation	Example sentence	type
MadeOf	NP is made of NP.	P
DefinedAs	NP is defined as NP.	P
HasA	NP has NP.	P
HasProperty	NP is AP.	P
UsedFor	NP is used for VP.	F
CapableOf	NP can VP.	F
HasPrerequisite	NP—VP requires NP—VP.	F
MotivatedByGoal	You would VP because you want VP.	F
...	...	...

is composed by a general cycle over all synsets in WordNet. Then, for each synset S_i we compute the set of all candidate semantic ConceptNet triples $P_{conceptnet}(S_i)$ as the union of the triples that contain at least one of the terms in T_i. The inner cycle iterates over the candidate triples to identify those that can enrich the synset under consideration. We used a number of heuristics to align each ConceptNet triple c_k (of the form $NP - rel - NP$) to each synset S_i:

h1 IF a lemma of an NP of the triple c_k is contained in the lemmatized gloss g_i of the synset S_i.This would mean that ConceptNet contains a relation between a term in T_i and a term in the description g_i, making explicit some semantics contained in the gloss. Note that the systematic inclusion of *related-to* relations with all the terms in the gloss g_i would carry to many incorrect enrichments, so an heuristic like **h1** is necessary to identify only correct enrichments.

h2 IF a lemma of an NP of c_k is also contained in $P_{wordnet}$. By traversing the WordNet structure, it is possible to link words of related synsets to S_i by exploiting existing semantics in ConceptNet.

h3 IF a lemma of an NP of c_k is contained in the lemmatized glosses of the most probable synsets associated to the words in g_i. The word sense disambiguation algorithm used for disambiguating the text of g_i is a simple match between the words in the triple with the words in the glosses. In case of empty intersections, the most frequent sense is selected.

h4 After taking all the hypernyms of S_i, we queried ConceptNet with their lemmas obtaining different sets of triples (one for each hypernym lemma). IF the final part $* - rel - word$ of the triple c_k is also contained in one of these sets, we then associate c_k to S_i. The idea is to intersect different sets of ambiguous common-sense knowledge to make a sort of collaborative filtering of the triples. For example, let S_i be $S_{burn,burning}$: *pain that feels hot as if it were on fire* and the two candidate ConceptNet triples $c_1 = burning - relatedto - suffer$ and $c_2 = burn - relatedto - melt$. Once retrieved *hypernyms*($S_{burn,burning}$) = $\{pain, hurting\}$ from WordNet, we query ConceptNet with both $pain$ and $hurting$, obtaining two resulting sets $P_{conceptnet}(pain)$ and $P_{conceptnet}(hurting)$. Given that the end of the candidate triple c_1 is contained in $P_{conceptnet}(pain)$, the triple is added to synset $S_{burn,burning}$. On the contrary, the triple c_2 is not added to $S_{burn,burning}$ since $relatedto - melt$ is not contained neither in $P_{conceptnet}(pain)$ and $P_{conceptnet}(hurting)$.

The proposed method was able to link (and disambiguate) a total of 98122 individual ConceptNet instances to 102055 WordNet synsets. Note that a single ConceptNet instance is sometimes mapped to more than one synset (e.g., the semantic relation *hasproperty-red* has been added to multiple synsets such as [*pomegranate, ...*] and [*pepper, ...*]). Therefore, the total number of ConceptNet-to-WordNet alignments was 582467. Note that we only kept those instances which were not present in WordNet (i.e., we removed redundant relations from the output). Table 4.4. shows an analytical overview of the resulting WordNet enrichment according to the used heuristics.

Heuristic	# of enrichments
h1	222544
h2	109212
h3	19769
h4	230942

Table 2: Overview of the WordNet enrichment according to the used heuristics.

In order to obtain a first and indicative evaluation of the approach, we manually annotated a set of 505 randomly-picked individual synset enrichments. In detail, given a random synset S_i which has been enriched with at least one ConceptNet triple $c_k =< NP - rel - NP >$, we verified the semantic correctness of c_k when added to the meaning expressed by S_i, considering the synonym words in T_i as well as its gloss g_i and examples E_i. Table 4.4. shows the results.

The manual validation revealed a high accuracy of the automatic enrichment. While the total accuracy is 88.31% (note that higher levels of accuracy are generally difficult to reach even by inter-annotation agreements), the extension seems to be highly accurate for relations such as *capable-of* and *has-property*. On the contrary, *is-a* and *related-to* relations have shown a lower performance. However, this

Relation	# correct	# incorr.	Acc.
related-to	121	22	84.62%
is-a	99	17	85.34%
at-location	39	5	88.84%
capable-of	36	1	97.29%
has-property	29	2	93.55%
antonym	27	4	87.10%
derived-from	25	1	96.15%
...	...	...	...
Total	446	59	88.31%

Table 3: Accuracy of some WordNet semantic enrichments obtained by the manual evaluation.

is in line with the type of used resources: on the one hand, WordNet represents a quite complete taxonomical structure of lexical entities; on the other hand, ConceptNet contains a very large semantic basis related to objects behaviours and properties. Finally, *related-to* relations are more easily identifiable through statistical analyisis of co-occurrences in large corpora and advanced topic modeling built on top of LSA (Dumais, 2004), LDA (Blei et al., 2003) and others. Extending WordNet with non-disambiguated common-sense knowledge may by challenging, also considering he very limited contextual information at disposal. However, such an alignment is feasible due to the few presence of common-sense knowledge related to very specific synsets / meanings (e.g., for the term "*cat*", it is very improbable to find a common-sense fact related to the synset S_{cat}: *a method of examining body organs (...)*.

5. Bibliographical References

Agirre, E., Ansa, O., Hovy, E., and Martínez, D. (2000). Enriching very large ontologies using the www. *arXiv preprint cs/0010026*.

Bentivogli, L. and Pianta, E. (2003). Beyond lexical units: Enriching wordnets with phrasets. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 67–70. Association for Computational Linguistics.

Bentivogli, L., Bocco, A., and Pianta, E. (2004). Archiwordnet: integrating wordnet with domain-specific knowledge. In *Proceedings of the 2nd International Global Wordnet Conference*, pages 39–47.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Chen, J. and Liu, J. (2011). Combining conceptnet and wordnet for word sense disambiguation. In *IJCNLP*, pages 686–694.

Collins, N. S. and Murphy, J. (2013). Towards a folk taxonomy of popular new media marketing terms.

Di Caro, L., Candan, K. S., and Sapino, M. L. (2011). Navigating within news collections using tag-flakes. *Journal of Visual Languages & Computing*, 22(2):120–139.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Hsu, M.-H., Tsai, M.-F., and Chen, H.-H. (2008). Combining wordnet and conceptnet for automatic query expansion: a learning approach. In *Information Retrieval Technology*, pages 213–224. Springer.

Kiu, C.-C. and Tsui, E. (2011). Taxofolk: A hybrid taxonomy–folksonomy structure for knowledge classification and navigation. *Expert Systems with Applications*, 38(5):6049–6058.

Laparra, E., Rigau, G., and Cuadros, M. (2010). Exploring the integration of wordnet and framenet. In *Proceedings of the 5th Global WordNet Conference (GWC 2010), Mumbai, India*.

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

Montoyo, A., Palomar, M., and Rigau, G. (2001). Wordnet enrichment with classification systems. In *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customisations Workshop.(NAACL-01) The Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 101–106.

Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F., and Cucchiarelli, R. (2004). Extending and enriching wordnet with ontolearn. In *Second Global WordNet Conference, Brno, Czech Republic, January*, pages 20–23.

Niemi, J., Lindén, K., Hyvärinen, M., et al. (2012). Using a bilingual resource to add synonyms to a wordnet. In *Proceedings of the Global Wordnet Conference*.

Niles, I. and Pease, A. (2003). Mapping wordnet to the sumo ontology. In *Proceedings of the ieee international knowledge engineering conference*, pages 23–26.

Pease, A., Niles, I., and Li, J. (2002). The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, volume 28.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2007). Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data & Knowledge Engineering*, 61(3):484–499.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., and Navigli, R. (2014). Validating and extending semantic knowledge bases using video games with a purpose. In *Proc. of ACL*.