Multilingual eXtended WordNet Knowledge Base: Semantic Parsing and Translation of Glosses

Tatiana Erekhinskaya, Meghana Satpute, Dan Moldovan

Human Language Technology Research Institute
The University of Texas at Dallas
{erekhinskaya, meghana, moldovan}@hlt.utdallas.edu

Abstract

This paper presents a method to create WordNet-like lexical resources for different languages. Instead of directly translating glosses from one language to another, we perform first semantic parsing of WordNet glosses and then translate the resulting semantic representation. The proposed approach simplifies the machine translation of the glosses. The approach provides ready to use semantic representation of glosses in target languages instead of just plain text.

Keywords: Knowledge base, semantic relation, synset, translation

1. Introduction

WordNet (Fellbaum, 1998) is one of the most popular and useful resource for the English language. Creation of such resources for other languages would make significant improvements in various NLP applications. However, manual creation is very expensive and might not be practical for less used languages. Automatic or semiautomatic translation can help with synset translation, but hardly can produce smooth glosses in target languages (Giménez and Màrquez, 2006), which means significant loss of knowledge. In addition, even if the translation of glosses results in absolutely correct text in the target language, plain text is almost impossible to use for languages without an advanced NLP toolkit. Our apprach provides ready to use semantic representation of glosses in target languages.

This paper offers a shortcut: translate the semantic structure of the glosses rather than glosses themselves. The idea is illustrated in Figure 1. The semantic structure can be extracted using the Polaris semantic parser (Moldovan and Blanco, 2012). The parser uses a fix set of 26 binary relations. For translation, the type of the relation remains the same, only arguments are translated. Since arguments are single words or (rarely) short phrases, this approach significantly simplifies the translation.

Several resources were obtained by this approach, namely Extended WordNet Knowledge Base containing semantically parsed WordNet glosses and its translations into Russian and Hindi[1]. Since translation is based on open resources Universal WordNet (de Melo and Weikum, 2009), Wikipedia and Wiktionary, the same approach can be applied for other languages.

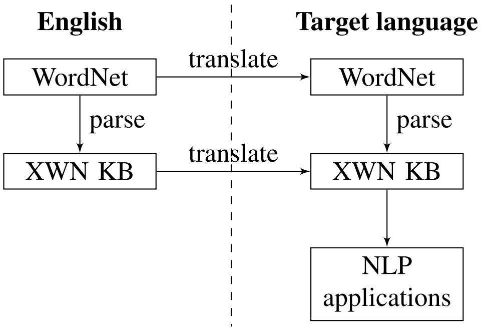

Figure 1: The shortcut idea.

2. Related Work

There are many WordNet-like resources for other languages, i.e EuroWordNet for some European languages[2], versions for Chinese, Japanese and Arabic languages. There is also Global WordNet Association[3] and AsianWordNet Project[4] facilitating the development of such resources.

For the Russian language there are several WordNet projects:

1. RussNet (Azarova, 2005) - manually populated resource, which allows to obtain thesaurus capturing specific features of Russian. It currently has about 5500 synsets: 2300 nouns, 1900 verbs, 1100 adjectives and 200 adverbs.

2. Russian WordNet (Balkova et al., 2004; Yablonsky, 2010; Yablonsky, 2013) semiautomatic creation of English-Russian WordNet 2.0 from

[1] Freely available at https://github.com/erekhinskaya/multilanguage-xwn-kb

[2] http://www.illc.uva.nl/EuroWordNet/

[3] http://www.globalwordnet.org/

[4] http://th.asianwordnet.org/

Linked Open Data resources WordNet 3.0, DBPedia and Yago2 (Hoffart et al., 2011). Available on http://www.tendrow.org/, crowdsourcing platform for manual reviewing and improving the results of automatic processing.

3. Wordnet.ru (Lypatov et al., 2005) - project aimed to automatically translate synsets. The authors were able to translate about 45% synsets, 75% of which correctly. However, their approach does not allow to get translations for glosses.

None of these resources offers both Inter-Lingual-Index and glosses in Russian.

The Center For Indian Language Technology (CFILT) at Indian Institute of Technology, Mubai, India is working on creating WordNet for different Indian languages similar to Princeton University's WordNet. The Hindi WordNet[5] (HWN) follows the design principles of the English WordNet while paying particular attention to language specific phenomena (such as complex predicates) whenever they arise. HWN has been created by manually looking up the various listed meanings of words in different dictionaries. The resource currently has 33900 synsets (Bhattacharyya, 2010).

3. Semantic Parsing of the Glosses

The first step of our approach is semantic parsing of glosses. We used Polaris semantic parser (Moldovan and Blanco, 2012). This parser uses a fixed relation set (Table 1) which was decided based on inventories used by others (PropBank, FrameNet, NomBank, SemEval competitions). Some relations considered elsewhere are ignored since they do not occur frequently enough in text and their automatic extraction would not be feasible, e.g., ENTAILMENT.

Most glosses do not have the structure of a full sentence, which makes semantic parsing harder. To solve the problem, we need to construct a smooth sentence from incomplete synset definitions typically contained in glosses. The input for the semantic parser is formed as:

⟨*first word from synset*⟩ + *"is"* + ⟨*gloss*⟩.

For example, for synset *pasha:n#1* we form an input string *'Pasha is a civil or military authority in Turkey or Egypt'*, and the Polaris output for the sentence is:

LOCATION(Turkey, authority)
LOCATION(Egypt, authority)
PART_WHOLE(authority, Turkey)
VALUE(civil, authority)
ISA(pasha, authority)

Relation	Example
CAUSE	CAU(earthquake, tsunami)
JUSTIFICATION	JST(it is forbidden, don't smoke)
INFLUENCE	IFL(missing classes, poor grade)
INTENT	INT(buy, Mary)
PURPOSE	PRP(storage, garage)
VALUE	VAL(blue, car)
SOURCE	SRC(Mexican, avocados)
AGENT	AGT(John, gave)
EXPERIENCER	EXP(John, felt)
INSTRUMENT	INS(a hammer, broke)
THEME	THM(flowers, gave)
TOPIC	TPC(issue, discuss)
STIMULUS	STI(train, listen)
ASSOCIATION	ASO(phone, fax)
KINSHIP	KIN(John, his cousin)
IS-A	ISA(convertible, car)
PART-WHOLE	PW(engine, car)
MAKE	MAK(BMW, cars)
POSSESSION	POS(John, truck)
MANNER	MNR(quick, delivery)
RECIPIENT	RCP(Mary, gave)
SYNONYMY	SYN(twelve, a dozen)
LOCATION	LOC(in the porch, gave)
TIME	TMP(yesterday, gave)
PROPERTY	PRO(height, John)
QUANTIFICATION	QNT(a dozen, roses)

Table 1: Inventory of relations extracted by Polaris (Moldovan and Blanco, 2012).

There should be also the relation VALUE(military, authority), which was not picked up by the semantic parser.

No existing semantic parser guarantees perfect quality of the result. Fortunately, both structures of glosses and errors of the semantic parser have some regularity that allows us to improve results using linear and syntax-based templates, similar to (Hearst, 1992).

We use several metrics to estimate quality of semantic parsing:

1. Connectivity is defined as a percentage of the glosses that contain at least one relation whose argument is the gloss synset. Such relations are very important, because otherwise other relations inside the gloss become unreachable from the gloss synset and therefore useless.

2. Coverage is defined as a percentage of concepts in glosses which are an argument of some relation.

3. Accuracy is defined as a percentage of correct relations in comparison with a gold standard.

The first two metrics can be measured automatically

[5] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

over the whole WordNet and give a kind of upper-bound estimation of the quality. The third metric can be estimated only in comparison with a manually annotated portion of the glosses.

Aiming to achieve a high degree of connectivity, we can connect explained words and the first word in the gloss in the following manner. For each POS of the explained synset, if the first word in the gloss has the same POS, we can connect them with an ISA relation. For adjectives, the gloss can be started with a noun (usually *something* or *someone*), then the relation VALUE should be added.

Similar rules can be used to improve coverage. A problematic structure for Polaris is the usage of *and/or* conjunctions. The typical error of the parser is that only one component is covered by some semantic relation, the other one is ignored. So, if there are two adjectives with a conjunction in between, and only one of them takes part in VALUE relation, we can add the missing relation for the other adjective. For example, for *Polar is something having a pair of equal and opposite charges* Polaris extracts VALUE(opposite, charges) but misses VALUE(equal, charges). The missing relation can be reconstructed, since *equal* and *opposite* both modify *charges*. A more complicated case is when two verbs are connected with a conjunction and share the same subject, i.e.: *Mess is a military dining room where service personnel eat or relax*. Since the second verb does not have its own subject, it is likely related to the subject of the first one. Thus, we can add AGENT(personnel, relax).

Another problem for Polaris is wh-clauses. It is possible to extract clauses as separate sentences, replace the pronouns with their antecedents and process the clauses with semantic parser to retrieve missing relations. For example, for the sentence *Sport is the occupation of athletes who compete for pay* only two relations were picked up initially: THEME(pay, compete) and ISA(sport, occupation). Extracting the clause and replacing the pronoun with the preceding noun, we get sentence *"athletes compete for pay"*, for which the semantic parser easily provides the AGENT(athletes, compete) relation.

4. Translation

After the semantic parsing of the glosses we perform word sense disambiguation of the relation arguments using Princeton University WSD annotation[6] for WordNet 3.0 glosses and mapped it back to WordNet 3.1[7]. We used several resources to translate the relation arguments represented as synsets into target language:

1. Universal WordNet[8] (de Melo and Weikum, 2009).

2. If Universal WordNet does not provide a translation, we use WordNet-Wikipedia mapping (Fernando and Stevenson, 2012; Fernando, 2013) to extract Wikipedia article title translated into a target language.

3. If WordNet-Wikipedia mapping does not provide a translation (which normally happens with POS other than nouns), we use translations extracted from Wiktionary.

This allows to get translations of semantic relations in glosses to Russian and Hindi. An example of such translation is shown in Table 2.

5. Evaluation

5.1. Semantic Parsing Gold Standard

To make a gold standard, 100 glosses were taken randomly for manual annotation, keeping the proportion of parts of speech equal to the one in WordNet. Examples with negations expressed by *not/no/without* or words like *cancel, fail* were eliminated. The annotators agreed on the following rules:

1. Compound nouns which are not WordNet concepts should be represented using VALUE, INSTRUMENT, PURPOSE and other relations. For example, *search system* is represented as PURPOSE(search, system).

2. Prepositions which do not add meaning to the relation should be eliminated, i.e *Price of a car* represented as PROPERTY(*price, car*). Prepositions giving additional information to the relation are kept: *created from nonliving matter* – MANNER(from matter, created).

3. We allow duality of PART_WHOLE and LOCATION: for some examples both PART_WHOLE and LOCATION make sense: *Car engine*, *city in the county*.

4. We allow subjects to be annotated as AGENT. Compare the following examples: *man drives*, *river flows*, *airplane flies*, *knife cuts*. The real AGENT is only the first phrase: subject is animated and makes action willingly. It is hard to

[6] http://wordnet.princeton.edu/glosstag.shtml

[7] http://www.hyperdic.net/en/doc/mapping

[8] http://www.mpi-inf.mpg.de/yago-naga/uwn/publications.html

English	Russian	Hindi
PART_WHOLE(clay, oven)	PART_WHOLE(глина, печь)	PART_WHOLE(माटी, भाठी)
LOCATION(Pakistan, use)	LOCATION(Пакистан, использовать)	LOCATION(पाक, उपयोग)
LOCATION(India, use)	LOCATION(Индия, использовать)	LOCATION(भारत, उपयोग)
VALUE(northern, India)	VALUE(северный, Индия)	VALUE(उत्‌तरी, भारत)
THEME(oven, use)	THEME(печь, использовать)	THEME(भाठी, उपयोग)
ISA(tandoor, oven)	ISA(тандыр, печь)	ISA(तंदूर, भाठी)

Table 2: An example of a translated gloss: *tandoor is a clay oven used in northern India and Pakistan.*

say, what *airplane* and *river* are, whereas *knife* ideally would be an INSTRUMENT. Since these cases are very subtle, agentive relation can be considered correct for any subject-action scheme.

5.2. Results

All glosses of WordNet 3.1 were processed with the Polaris semantic parser. We estimated the results using the connectivity, coverage and accuracy defined in Section 3. The accuracy is 69% for the output of the Polars parser and 74% after applying correction patterns. The values of connectivity and coverage are given in Table 3 below. The statistics of collected semantic relations is given in Table 4. For relations not mentioned in the table, i.e INFLUENCE, no instance was extracted from the glosses.

	Noun	Verb	Adjective	Adverb
Connectivity	85 / 90	57 / 59	51 / 55	14 / 19
Coverage	69 / 78	62 / 64	59 / 63	47 / 50

Table 3: Quality of semantic parsing before / after the application of correction patterns.

Relation	Noun	Verb	Adjective	Adverb
ISA	90194	6551	9751	164
POS	7444	151	631	13
PW	42900	638	3384	92
AGT	19062	823	5382	27
SYN	91	-	-	-
KIN	636	3	10	-
QNT	21716	1555	3671	214
VAL	95435	6843	18265	2483
THM	52633	6403	8051	223
LOC	31942	939	2904	50
MNR	8270	2173	3996	248
CAU	535	46	123	1
TMP	5485	241	701	179

Table 4: Statistics of relations in semantically parsed glosses.

The semantic structures of the glosses were translated into Russian and Hindi. For the Russian language we managed to translate about 80% of relation arguments, 64% of them correctly. For Hindi we translated 56% of relation arguments, about 68% of them correctly. Universal WordNet provides about 47% of translations, Wikipedia about 37% and Wiktionary 16%.

5.3. Error Analysis

We randomly picked 1000 glosses keeping the POS proportion in the subset. The reviewing revealed the following typical errors:

1. Missing relation between word from synset and main word in the gloss. The difficult situation for semantic parsing is when an adverb is explained with a prepositional phrase, like for example *"much"* – *"to a great degree or extent"* or when an explanation is rather grammatical than semantic: *"very"* – *"used as intensifiers"*.

2. Glosses with homogeneous constituents connected with conjunctions and/or with intermediate words. For example, for the sentence *Somber is something grave or even gloomy in character* the relation VALUE(grave, character) is missing. It can not be corrected by patterns because of the word *even*.

3. Mislabeled QNT relation. For example, for *"pitchy is something having the characteristics of pitch or tar"* the output relations are mislabeled:
   QUANTITY(characteristics, pitch)
   PART_WHOLE(characteristics, tar)
   ISA(pitchy, something).

6. Discussion

WordNet glosses contain many relations but encoded in plain text. The semantic parsing of the glosses extracts these relations and also provides an easy way to translate glosses into other languages.

The representation based on a fixed set of 26 relations allows consistency of the representation: all words from the text are represented as arguments of the relations and can be processed in a uniform manner. The approach proposed in the paper can be easily extended to other low-resource languages.

7. References

I. V. Azarova. 2005. RussNet project. `http://project.phil.spbu.ru/RussNet/`.

V. Balkova, A. Sukhonogov, and S. Yablonsky. 2004. Russian WordNet. From UML-notation to Internet/Intranet Database Implementation. In *Proceedings of the Second International WordNet Conference*, Brno, Czech Republic.

P. Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

G. de Melo and G. Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

S. Fernando and M. Stevenson. 2012. Mapping wordnet synsets to wikipedia articles. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

S. Fernando. 2013. Enriching lexical knowledge bases with encyclopedic relations.

J. Giménez and L. Màrquez. 2006. Low-cost enrichment of Spanish WordNet with automatically translated glosses: combining general and specialized models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 287–294, Sydney, Australia, July. Association for Computational Linguistics.

M. A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the COLING 1992*.

J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum. 2011. Yago2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 229–232, New York, NY, USA. ACM.

A. Lypatov, A. Goncharuk, I. Gelfenbein, V. Shylo, and V. Lehelt. 2005. Russian WordNet. `http://wordnet.ru/`.

D. Moldovan and E. Blanco. 2012. Polaris: Lymba's Semantic Parser. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

S. Yablonsky. 2010. English-Russian WordNet for Multilingual Mappings. In *Proceedings of 2010 Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web*, Shanghai, China.

S. Yablonsky. 2013. Web 2.0/3.0 Technology in Lexical Ontology Development: English-Russian WordNet 2.0. In *Proceedings of 4th International Conference Knowledge Engineering and the Semantic Web*, St. Petersburg, Russia.