Improving Automatic Ontology Developement

Marek Grác, Adam Rambousek

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
xgrac@fi.muni.cz, xrambous@fi.muni.cz

Abstract. This article describes the approach to build a new semantic network, which contains not only positive semantic labeling, but also the negative information. In order to obtain high quality data for the following use in machine learning and machine translation, we have created method based on automatically pre-generated data from the large corpora, followed by manual annotation. In this way, the core of semantic network was produced, which can be expanded to improve corpora coverage.

Key words: semantic network, ontology, ontology development

1 Introduction

To improve current complex NLP application we need to explore semantic level of natural languages. Both prevalent techniques (rule based and machine learning) tend to use information obtained from morphology and syntax level what lead to a situation where we cannot expect major improvement without adding something new. It can be completely new algorithms, more computing power or new data resources. In this paper we would like to focus on creating a new machine readable semantic network that will fill an existing gap.

Semantic networks and ontologies are used in NLP for several last decades. These networks focus on various aspects of semantic layer. There are ones which cover word senses (e.g. WordNet [1]) and other that are on the boundary of semantic and pragmatic layer of language (e.g. CyC [2]). The focus on semantic network is heavilly dependent on target applications. Ours goals are improving of syntactic parsing and as a result improve information extraction process from free text. Words in a language are very often bound to a specific set of words in language. These relations are traditionally called valencies.

Valencies, in various forms, are present for almost every PoS in language. The most studied ones are traditionally verb valencies. We have several verb valencies lexicons based on different linguistic theories which targets different usage. The most known are VerbNet [3], FrameNet [4] and Pattern Dictionary of English Verbs [5]. It is very important that such resources are consistent, they have acceptable coverage and that they are in superior quality. Such high expectations means that their development is expensive and unobtainable for smaller languages. Automatic methods of creating verb valencies using

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 105–110, 2012.

unsupervised methods of machine learning on unlabelled corpora are also available. Their quality vary a lot depending on language and used resources but in general hand-made (or semi-automatic) methods are still better even if they have lower coverage on text in corpora.

If we can automatically obtain valencies then we can create a thesaurus or even semantic network directly. But we can do it also in reverse order. If we have semantically labeled corpora obtaining a valency lexicon is no longer such difficult problem. In this paper we will show that we can iterativelly detect valencies and improve semantic network. In current state we cannot create valencies for verbs but we will use methods which are simple but suprisingly precise.

2 Semantic Network

Existing shallow semantic networks are influenced by Princenton's WordNet [1]. In fact WordNet is probably only network which lead to development of similar projects, for example different languages like EuroWordNet [6] and BalkaNet [7], or extensions like OntoWordNet [8] and eXtended WordNet [9].

For our experiment we have decided to use a Czech language because there is a Czech wordnet and thus ours results can be easilly compared to existing resources. As we expect to use proposed methods for languages without deep NLP support we will use only limited amount of technologies. Czech language is highly inflective with (almost) free word order in sentence. This means that direct usage of statistical methods will not work perfectly due to sparsness of data. For this purpose we will use morphological analyzer [10], guesser for unknown words and tagger / lemmatizer to obtain corpora with desambiguated morphological information. Used tagger 'desamb' [11] has precision sligtly over 90% and it's precision is very close to precision of taggers for other smaller languages. Existing Czech WordNet has almost same structure as Princeton one and we will use english names of elements across this paper.

When we take a look at verb valency dictionaries then the semantic class which is one of the most common is 'person' which is usually in position of 'agens'. In Czech language position of subject is in specific case 'nominative' but word-form representing this case is systematically same with case 'accusative' which represents object. Due to free word order, we are unable to obtain very high precision in this situation with just morpho-syntactic information.

Position of 'subject' with semantic class 'person' is very common but only very rarely subject is bound only to this specific class. More often there are also other semantic classes: institution and animals. These three classes are used in very similar way and it is very difficult to distinguish person and institution.

John loves Apple.
John loves Mary.

Simplest solution is to to create a new semantic class which will contain all such classes. Then we are in situation when John, Apple or bakery are in a same class because these words can be represented by person (or animal).

Bakery should also be in semantic class 'location' or 'building'. In WordNet-like networks this is done by adding new 'sense' of word which is hyponym for 'building'. We prefer not to fragment senses of words into separate categories and we do not target to an application that can uses such information. This was a main reason why we have decided to just add attributes to words. These attributes can be in hyponymy/hyperonymy relations.

We also believe that static (hard-coded) language resources are incomplete and they cannot be completed for living languge. This lead us to prefer open world assumptions (OWA) [12] in our ontology. OWA means that information which cannot be disapproved can be valid. Missing words in semantic network can have any available attributes. Because of OWA we have decided to populate the semantic network not only with positive information (John can have attribute *freewill*) but also negative information (table cannot have attribute *freewill*). Using such negative information helps us to properly work with words and valencies when we do not know what it is but we know that this can't be 'person' for example. In fact our preliminary results show us that these negative information are more usefull for syntactic parsing than positive ones. Mainly because quality of syntactic parser is so high that more often we will just confirm correct answer by semantic network but negative information will show us possible errors.

Problem with attributes instead of direct usage of hyponymy/hyperonymy relations is that we (in ideal world) have to work with every word in language. Expenses for annotation are then quite linear. For N semantic classes we have to answer N question for every word. Hypo/hypero relation between attributes can helps us to have N sufficiently low.

Annotation framework SySel [13] can be used to distribute and work with dividing words into categories. For yes/no questions the average answer is gathered from annotator in 1-2 seconds what lead to approx. 2,000 words / hour. Even if we need to annotate each word by several annotators, this process is really fast for smaller groups of words / semantic classes. If we need to distinguish attributes like can-it-looks-like-pumpkin? then it is very effecient way. Even if we have to handle usually only tens of thousands words (in our experiment approx. 100,000) then we would like to improve possibility to automatically add new words into semantic classes.

3 Extending existing semantic network

At the start of our experiment we did not focus on automatic methods of extending semantic network. Our decision was to create a completely new and free semantic network which will improve our existing tools. Creating a semantic network was not a goal of this process and that is main reason why our network has still huge gaps in semantic classes. We prefer do create a new semantic classes in a momement when we expect that they will improve tools not because we wish that they will maybe help one day.

In the first stage of project, 30 student annotators works with SySel to annotate 100,000 words if they can be *freewill* or they can't be. Each word was annotated at least two times and to accept a word/attribution/boolean, the metrics showed in table 1 were used.

Table 1. Accept metrics

# of annotators	non-agreement-annotation
2 - 4	0
5 - 8	1
9 - 12	2
12 +	3

We were able to create a semantic network that consists of 4,538 words which have attribute *freewill* and 101,303 that cannot have attribute *freewill*. More than 9,000 words didn't have the annotator agreement high enough to add them to semantic network. Relatively high level of inter-annotator error was probably due the fact of using work-power where some students did not focus on work (e.g. man is not *freewill* even if word is in examples) and only partially due to borderline words (e.g. democracy) that were not specified in annotation manual.

Semantic network have to be incomplete but we can attempt to improve it and extend it. We decided to test the most simple options of using existing semantic network and parsed corpora to do it.

In most of the language with free word order it is very difficult to match subject/verb and similar relations and full syntactic analysis is needed. But usually there are at least some rules that works pretty well. For Czech language we were able to identify those three:

- preposition, noun
- adjective, noun (if they have agreement in case, gender, number)
- noun, noun in genitive case - construction similar to english 'X of Y'

Very fast we found out that there is no preposition which is bound exclusively with *freewill*. Number of adjectives and nouns lead us to develop an automatic finder for such situations.

We want to find such words that are bound (almost) exclusively with given semantic class *freewill* using existing semantic network. From parsed corpora we will extract all bigrams which match our morphological rules. Then we will prepare stastistic for usage of each adjective with semantic class. These statistics is later filtered to contain only those adjectives which are used (almost) exclusively with words with possitive attribute *freewill*. Words which misses that attribute in semantic network (or they are not in network at all) will be accepted if there are enough adjectives that are used together with this word.

What are the main problems of this process?

- Our attributes means that word represents this attribute in one of its senses, this is important because in some cases we can detect adjective together with word-sense which we do not plan. e.g five-pointed star vs rock star.
- We can find out adjectives which are only partially relevant. Quite a lot of found adjective belongs to group of adjective that represents 'X years old'. Our system correctly does not find one-year old, five-years old because there are lot of mentions of wine, whiskey, buildings, ... and it will correctly find 72-years old as such numbers are usually specific for person. Very similar process is in place for possesive adjective (e.g. dog's, Achilles's).
- As we are working with bigrams directly it is possible that we will add a word which do not have semantic representation of attribute directly. e.g He ate a "box" of chocolate. The word 'box' works just as a modifier in this context. We can detect these words because they will occur as positive examples for most of the attributes.

Process itself is relatively fast and on 350 million corpora it took less than 3 hours to make one iteration. Due to quality of data we can add found words to semantic network automatically and re-run the process. Our experiments showed that few iteration will drastically improve coverage and this method can very easilly solve border-line cases where human annotators are not sure. Border-line cases does not have to be solved consistently but we can add words only to positive side of semantic network.

tabulka s vysledkami

Table 2. Coverage

	# of words identified	k1 coverage	'k2 k1' coverage
manual	105,840	68.26%	94.28%
after 1st iteration	106,044	74.48%	96.08%
after 2nd iteration + selected proper nouns[1]	106,250	81.49%	97.99%
after 3rd iteration	106,942	83.07%	98.6%

Table 3. Random samples of 10,000 words with *freewill* attribute, called seed1, seed2, seed3

sample	k1 coverage	new words precision
seed1, iteration 2	25.51%	84.50% (780 words)
seed2, iteration 2	40.76%	75.78% (1514 words)
seed2, iteration 3	33.19%	72.24% (788 words)

4 Evaluation and future extensions

As seen in table 3, automatic detection algorithm combining manual annotation of small starting set, valency detection, and limiting linguistic rules works very well in identifying word attributes, ie. can say that word belongs to particular semantic class.

Negative attributes, ie. information that word does not belong to semantic class, are very useful feature for the applications using the semantic network. Such knowledge can reduce the time needed to parse the text, for example. However, negative attributes in our semantic network are annotated manually only. Current rules and tools for automatic annotation does not provide precision good enough to include in semantic network. Improvement of negative attributes detection is one of the next steps of this project. Coverage enhancement is the other big goal, both in terms of annotated words, and different semantic classes.

Acknowledgements This work has been partially supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by EC FP7 project ICT-248307.

References

1. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998)
2. Lenat, D., Guha, R., Pittman, K., Pratt, D., Shepherd, M.: Cyc: toward programs with common sense. Communications of the ACM **33**(8) (1990) 30–49
3. Schuler, K.: VerbNet: A broad-coverage, comprehensive verb lexicon. PhD thesis, University of Pennsylvania (2005)
4. Fillmore, C., Baker, C., Sato, H.: Framenet as a 'net'. In: Proceedings of Language Resources and Evaluation Conference (LREC 04). Volume vol. 4, 1091-1094., Lisbon, ELRA (2004)
5. Hanks, P., Pustejovsky, J.: A pattern dictionary for natural language processing. Revue Française de linguistique appliquée **10**(2) (2005) 63–82
6. Vossen, P., ed.: EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer (1998)
7. Christodoulakis, D.: Balkanet Final Report, University of Patras, DBLAB (2004) No. IST-2000-29388.
8. Gangemi, A., Navigli, R., Velardi, P.: The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE (2003) 820–838
9. Mihalcea, R., Moldovan, D.: extended wordnet: Progress report. In: Proceedings of NAACL Workshop on WordNet and Other Lexical Resources. (2001) 95–100
10. Šmerk, P.: Fast morphological analysis of czech. RASLAN 2009 Recent Advances in Slavonic Natural Language Processing (2009) 13
11. Šmerk, P.: K morfologické desambiguaci češtiny. (2008)
12. Reiter, R.: On closed world data bases. Technical report, Vancouver, BC, Canada, Canada (1977)
13. Grác, M., Rambousek, A.: Low-cost ontology development. (In: 6th International Global Wordnet Conference Proceedings) 299–304