Language Technology: Research and Development - Project Proposal

Arvid Lindahl

October 21, 2014
1 Purpose and aims

The general purpose of this research project is to further the knowledge of multiword expressions within the field of language technology. More specifically, this project will look at means to describe the semantic relation holding between the two nouns in a two-word noun compound (NC). Recent research has seen interest in describing these relations by using paraphrases of the NCs and in particular by looking at verbs in these paraphrases. Although experiments with other types of paraphrases have been made, no comparison between the different types of paraphrases has been performed. Therefore, the goal for this project is to investigate what different types of paraphrases there are that can be used to describe the semantic relations of NCs, and also to evaluate these types to find out which ones are best suited for this task. The hope is that this project will show the usefulness of various types of paraphrases.

2 Survey of the field

Much of the research on semantic interpretation of NCs within language technology has focused on describing the semantics as a relation holding between the constituents in the NC. In particular, focus has primarily been on two-word NCs. This is due to the recursiveness of NCs: the same semantic relations are assumed to hold also for NCs longer than two words (Girju et al., 2005).

Both within theoretical linguistics (Levi, 1978; Warren, 1978; Lauer, 1995) and language technology (Moldovan et al, 2004; Kim and Baldwin, 2006; Girju, 2007) the idea of describing the semantics of NCs according to a finite set of relations has been popular. However, some linguists have argued that that the set of relations that can hold between the nouns in NCs are (almost) infinite (Downing, 1977). The idea of a more fine-grained approach to describing semantic relations between NCs has in recent years become popular within the field of language technology as well. One method to achieve this more fine-grained description that has been used is by looking at paraphrases of NCs. Take for example the NC cancer doctor. It can be paraphrased as a doctor that treats cancer. The verb treat(s) is then used as a way to describe the relation that holds between the nouns. Using paraphrasing verbs in this manner has been tested by Nakov (2008) and there has also been a shared task devoted to verbal and prepositional paraphrases (Butnariu et al., 2010). Nakov and Hearst (2013) applied this idea to build a verb-based vector space model for a given noun compound. They note, however, that there are many other ways of paraphrasing NCs than using verbs. In a SemEval-2013 shared task (Hendrickx et al., 2013) the use of free paraphrases of NCs was explored. Here, not only verbs or prepositions in the paraphrases were considered but the task allowed for paraphrases of various complexity.
3 Programme description

The research in this project intends to build on previous research on semantic interpretation of NCs. Nakov and Hearst (2013) presented a vector space model of verbs from paraphrases extracted from the web using a pre-defined search pattern. Free paraphrases were examined in a shared task in SemEval-2013 (Hendrickx et al., 2013). This project intends to map the different kinds of paraphrases that exist, construct search patterns that will allow for extraction of the key components in these paraphrase types, build vector space models for NCs using the different components of the paraphrases types, and, finally, evaluate the paraphrase types according to how well they can be used to describe the semantic relations of NCs.

The first step in this project is to investigate what different types of paraphrases there are. The intention here is to make use of the paraphrases gathered for the SemEval-2013 task on free paraphrases and define the types after studying this dataset. After this list of paraphrase types has been formulated, search patterns to capture the different types will be devised. Then, these search patterns can be used to extract paraphrases for NCs.

The test data and evaluation method have yet to be decided on. Since no similar experiment has been performed there is no standard evaluation method to follow. Nakov and Hearst (2013) perform both intrinsic and extrinsic evaluation in their paper. They compare extracted paraphrases of NCs to abstract relation sets used in previous studies and to paraphrases formulated by human annotators. They also evaluate what effect paraphrasing has in a machine translation system. However, it is difficult to apply these evaluation methods for this project. The point of the project is to determine how well different paraphrase types can capture the semantics of NCs. The evaluation method would have to be something that can capture this.

<table>
<thead>
<tr>
<th>Week</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>studying NCs, defining paraphrase types</td>
</tr>
<tr>
<td>45</td>
<td>defining search patterns/evaluation</td>
</tr>
<tr>
<td>46</td>
<td>evaluation</td>
</tr>
<tr>
<td>47</td>
<td>evaluation/paper</td>
</tr>
<tr>
<td>48</td>
<td>evaluation/paper</td>
</tr>
<tr>
<td>49</td>
<td>paper</td>
</tr>
<tr>
<td>50</td>
<td>paper</td>
</tr>
</tbody>
</table>

**Table 1:** Timeplan for the project

References

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2010. SemEval-2010 Task 9: The interpretation of


