Paraphrases in LFG-based Broad-Coverage Semantics

Sina Zarrieß and Jonas Kuhn
Linguistics Department, University of Potsdam, Germany

This paper presents an approach to the integration of paraphrases in an LFG-based, semantic processing system and outlines advantages of deep linguistic processing for dealing with paraphrases in large-scale semantics. We describe an XLE-based conversion routine that integrates the paraphrases acquired in a parallel corpus into the semantic system, providing a simple and effective representation for the meaning of phrasal expressions.

Our paraphrase representation extends the semantic transfer system outlined in (Crouch and King, 2006) which makes use of XLE's term rewrite engine to derive semantic representations from LFG F-structures. The system already integrates various lexical semantic resources like WordNet such that, e.g., (i) lexical items are linked to their synonyms, (ii) for diathesis alternations, verb arguments are captured in a canonicalized form, based on their thematic role, and (iii) regular processes of derivational morphology feed a canonicalized representation (Crouch and King, 2005). The resulting augmented and canonicalized meaning representation has the advantage that it can be directly used in processing tasks like Question Answering and Recognizing Textual Entailment: no inference module is required (thus avoiding hard computational problems and the resource bottleneck for inference rules), as semantic equivalence is represented explicitly and can be exploited by a comparatively simple matching strategy.

Of course, the effectiveness of this strategy of "deductive closure" over meaning representations depends on the quality, depth and coverage of the (near-)equivalence operations captured during analysis. Here, a considerable limitation of the currently available lexical resources is that they mostly encode semantic information about lexical items on the word-level. For large-scale semantic applications, it would be highly desirable to also detect semantic relations like synonymy between phrasal expressions. For instance, the following pair of sentences should be assigned the same meaning by relating the complex expression *put obstacles in the way of* to the atomic semantic relation *impede*.

(1) The European Union puts obstacles in the way of importing genetically-modified products.

(2) The European Union impedes the import of genetically-modified products.

A major challenge for paraphrase detection and representation is the large flexibility in their surface realizations. For instance, the paraphrase in (1) could be realized by modifying or replacing its components with synonyms such that it still conveys the meaning of *impede* like in (3).

(3) The European Union creates an obstacle to the import of genetically-modified products.

This variance in the surface realizations of paraphrases and the resulting low type frequencies complicate data-driven paraphrase acquisition substantially. Therefore, our approach for modelling phrasal semantic correspondences is based on a combination of shallow and deep linguistic processing techniques. (i) To assure a reasonable recall, a reliable paraphrase detection technique should operate on a large-scale resource. Our paraphrase extraction is based on the large parallel Europarl corpus and exploits shallow dependency analyses in addition to word alignments. (ii) We argue that the representation of the detected paraphrases should abstract as much as possible from their surface realization. While this is hard for surface-oriented corpus extraction approaches to achieve, it is rather straightforward in an LFG setting. In the framework of Crouch and King (2006), we can map complex phrasal expressions that may relate argument slots at various levels in a hierarchical embedding structure to a simple semantic predicate with corresponding argument slots.

Our proposal includes a routine for generating such transfer rules (which do get unwieldy for larger phrasal units) automatically from text instances, exploiting the XLE parsing system.

The **paraphrase detection** technique we propose combines ideas from graph-based monolingual approaches (e.g. (Lin and Pantel, 2001)) and alignment-based multilingual approaches (e.g. (Bannard and Callison-Burch, 2005)), drawing on structural and distributional information at the same time. The main idea is that if some boundaries of a given dependency configuration are perfectly aligned, but their intervening heads are not, it is very likely that this intervening syntactic material however corresponds to each other. The idea is illustrated in figure 1. In this paper, this strategy is explored for English and German verb alignments where the verbs arguments are taken as the boundaries of the dependency configurations.

The Europarl dependency parses which constitute the basis for paraphrase acquisition have been obtained by means of the MaltParser and stored in form of a relational database which can be efficiently

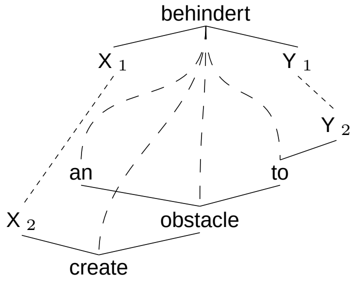

Figure 1: Example of a paraphrase configuration

queried. The detection procedure makes use of a generate-and-filter strategy where the candidate search is implemented as a dependency configuration search on the database. Then, a series of manually designed, linguistically motivated filters which mainly impose some rough conditions on the properties of the dependency path is applied to the candidate set to filter noise which is due to alignment and parsing errors. A preliminary evaluation on a testset of manually corrected alignments yields around 60% precision and 24% recall where word alignment yields around 20% precision and recall.

This extraction module outputs corpus instances of paraphrases as bilingual sentence pairs that exhibit an aligned dependency configuration as illustrated above. In the **conversion routine** these bilingual instances are mapped to a monolingual semantic correspondence and converted to a set of **generalizing transfer rules** which abstract away from all instance-specific aspects (such as the concrete fillers of the argument slots that do not form part of the paraphrase content) and can be straightforwardly used as a lexical knowledge base for other transfer-based semantic applications like PARCs ACQUAINT QA-system (Bobrow et al., 2007).

Our automatically generated transfer rules map the semantic abstraction of the paraphrase realization (corresponding to the left side of a transfer rule) to the semantics of its atomic realization (corresponding to the right side of the transfer rule). Given an atomic semantic lexical item in a source language, the left and right rule sides can be separately generated: the most frequent, atomic translation of the source item will figure on the right side of the transfer rules for each of the paraphrases of a given source item. The respective rule sides are obtained by using XLE to parse the target sentences, mapping them onto a semantic representation and stripping a set of pre-defined, semantic clauses irrelevant to the semantics of the paraphrase like tense or number. The dependency boundaries can then be replaced by variables on both sides of the rule.

```
+context_head(%t,put:%put),
+in_context(%t,role(sem_obj,put:%p,obstacle:%o)),
+in_context(%t,role(sem_subj,put:%p,%X)),
+in_context(%t,role(prep(of),way:%w,%Y)),
+in_context(%t,role(sem_obl(in),put:%p,way:%w))
==>
context_head(%t,hinder:%p),
skolem_info(hinder:%p,hinder,verb,verb,%p,%t),
in_context(%t,role(sem_subj,hinder:%p,%X)),
in_context(%t,role(sem_obj,hinder:%p,%Y)).
```

Figure 2: Example of a paraphrase representation as a transfer rule.

As an example, consider the transfer rule generated for the paraphrase *X put obstacles in the way of Y* detected as an alignment of the German verb *behindern* illustrated in figure 2. Since the English verb *X hinder Y* was found as the most frequent atomic translation, it constitutes the right rule side such that the subject of the paraphrase corresponds to the subject of *hinder* and the item embedded in the PP corresponds to the object of *hinder*. Since in the generation step, semantic features of the paraphrase candidate sentence related to the specific realization in context are eliminated, the transfer rule will fire on a set of sentences much larger than the set detected by surface overlap.

References

Bannard, Colin, and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 597–604, Morristown, NJ, USA. Association for Computational Linguistics.

Bobrow, Daniel G., Bob Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, Valeria de Paiva, Charlotte Price, and Annie Zaenen. 2007. PARC's Bridge question answering system. In Tracy Holloway King and Emily M. Bender (eds.), *Proceedings of the GEAF (Grammar Engineering Across Frameworks) 2007 Workshop*, pp. 13–15.

Crouch, Richard, and Tracy Holloway King. 2005. Unifying lexical resources. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pp. 32–37.

Crouch, Richard, and Tracy Holloway King. 2006. Semantics via F-Structure Rewriting. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG06 Conference*.

Dyvik, Helge. 2004. Translations as semantic mirrors. From parallel corpus to WordNet. *Language and Computers* 1:311 – 326.

Lin, Dekang, and Patrick Pantel. 2001. Dirt - discovery of inference rules from text. In *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 323–328.

Pado, Sebastian, and Mirella Lapata. 2005. Cross-lingual projection of role-semantic information. In *Proceedings of HLT/EMNLP 2005*, Vancouver, BC.