Fine-grained valence acquisition from large corpora for treebank grammars

Tejaswini Deoskar
School of Informatics
University of Edinburgh

Corpora with detailed annotations of linguistic structure, such as treebanks, are necessary resources for many aspects of language research but are expensive to create and limited in size. Due to the Zipfian nature of language, many linguistic events, especially those related to words, are inadequately represented in even large treebanks. For instance, fine-grained valence information obtained from a treebank is limited to a small set of high-frequency verbs, with most verb types in the language being either unseen or having too few occurrences in the treebank to be useful for statistical purposes. In this work, we describe a framework to enhance lexical information present in a treebank by capturing fine-grained information from unlabeled (raw) corpora. The framework has two aspects: one is the development of an appropriate grammar based on the treebank, that allows us to represent linguistically interesting properties. The second aspect is learning accurate statistical distributions from unlabeled data, a difficult task for fine-grained valence categories, given the ambiguity in language. These are described in turn below. The framework can be used for the purpose of valence acquisition for languages that have existing treebanks in the annotation style of the Penn Treebank II (for English) (Marcus et al., 1993).

Standard treebanks that exist or are being created for many languages consist of labeled trees with indexed empty categories. The choice of a fairly simple formal vocabulary ensures ease and consistency of annotation, and has led to significant advances in statistical parsing. But for other purposes (including linguistic research, lexicographic research and development, and research on sophisticated parsing models), it is a drawback that features such as inflectional categories, lemmas, subclassification of clausal categories, valence, and localized information about long distance dependencies are not overtly available in treebanks, although this information may be present in a non-overt form. We describe a methodology for augmenting a treebank database with such features, focusing our development effort on valence features and features which are relevant to licensing empty categories. The methodology has been applied to the Penn Treebank (Marcus et al., 1993) for English (Deoskar and Rooth, 2008) but can potentially be used with treebanks in other languages, with most software components being built to be reusable. A statistical grammar that be used with a parser is then extracted from the augmented treebank. We work with grammars which contain valence features on the pre-terminal symbols of verbs, with a fine-grained valence vocabulary (for instance, say, distinguishing verbs with subcategorized prepositions (*rid their beaches of medical waste*) from small clause verbs (*want Mr. Lawson in the office*). The framework is designed to allow easy manipulation of the valence vocabulary by the linguist. For instance, one may chose to lexicalise the preposition in a prepositional valence, or add other valence-related features

such as features classifying subjects (e.g. expletive subjects), and compile a grammar accordingly.

A treebank grammar and lexicon with a fine-grained valence vocabulary may not be useful by itself, due to the lexical sparsity of a treebank of even a large size. This lexical sparsity can be addressed only by moving to unlabeled data. In order to learn fine-grained but ambiguous valence information from unlabeled data, we use techniques based on the Expectation Maximization learning algorithm (Dempster et al., 1977), which has been shown in previous research to give positive results for valence acquisition, for both treebank grammars (Deoskar, 2008) and hand-crafted grammars (im Walde, 2002). For a current valence vocabulary size of 81 categories, we obtain substantial improvements (up to 17.5% error reduction over a smoothed treebank baseline) for verbs that are either unseen or have too few occurrences in the treebank to learn an accurate valence distribution. The fine-grained and broad-coverage valence lexicons thus obtained are aligned to the treebank database and the treebank grammar: this is an advantage because it allows direct incorporation of the lexicon in a treebank-based parser, plus objective evaluations and interpretations based on the treebank standard. In this respect, the framework differs from much previous work on valence acquisition from unlabeled data (Korhonen, 2002; Przepiórkowski, 2009), where either a shallow or deep parser (either treebank-based or not) is used to parse unlabeled data, but there is no direct relation between the valence vocabulary and the parser's grammatical representations. In addition, the framework allows the linguist to propose and include features of interest in the lexical representations. We are currently experimenting with a more fine-grained valence vocabularysuch as distinctions between different wh-complements. The framework thus allows a linguist to experiment with different features by incorporating them in lexical representations and to learn statistical distributions for them from large data, with a potential to inform linguistic research.

References

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Royal Statistical Society*, 39(B):1–38.

Tejaswini Deoskar. 2008. Re-estimation of Lexical Parameters for Treebank PCFGs. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*.

Tejaswini Deoskar and Mats Rooth. 2008. Induction of Treebank-Aligned Lexical Resources. In *Proceedings of 6th LREC, Marrakech, Morocco*.

Sabine Schulte im Walde. 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*.

Anna Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. thesis, Univ. of Cambridge.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Adam Przepiórkowski. 2009. Towards the Automatic Acquisition of a Valence Dictionary for Polish. In Magorzata Marciniak and Agnieszka Mykowiecka, eds., *Aspects of Natural Language Processing*, vol. 5070, pages 191–210. Springer-Verlag, Berlin.