High Quality Dependency Selection from Automatic Parses

Gongye Jin, Daisuke Kawahara, Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
jin@nlp.ist.i.kyoto-u.ac.jp {dk,kuro}@i.kyoto-u.ac.jp

Abstract

Many NLP tasks such as question answering and knowledge acquisition are tightly dependent on dependency parsing. Dependency parsing accuracy is always decisive for the performance of subsequent tasks. Therefore, reducing dependency parsing errors or selecting high quality dependencies is a primary issue. In this paper, we present a supervised approach for automatically selecting high quality dependencies from automatic parses. Experimental results on three different languages show that our approach can effectively select high quality dependencies from the result analyzed by a dependency parser.

1 Introduction

Knowledge acquisition from a large corpus has been actively studied recently. Knowledge is often acquired from the fundamental analysis. In particular, dependency parsing has been used for some tasks like case frame compilation (Kawahara and Kurohashi, 2006), relation extraction (Saeger et al., 2011) and paraphrase acquisition (Hashimoto et al., 2011). For these tasks, the accuracy of dependency parsing is vital. Although the accuracy of state-of-the-art dependency parsers for some languages like English or Japanese is over 90%, it is still not high enough to acquire accurate knowledge, not to mention those difficult-to-analyze languages like Chinese and Arabic.

Instead of using all the automatic parses, it is possible to use only high quality dependencies for knowledge acquisition. In this paper, we present a supervised approach for selecting high quality dependencies from automatic dependency parses. This method considers linguistic features that are related to the difficulty of dependency parsing. The experimental results on English, Chinese and Japanese show that our proposed method can select dependencies of higher quality than baseline methods for all the languages.

2 Related Word

There have been a few approaches devoted to automatic selection of high quality parses or dependencies. According to selection algorithms, they can be categorized into supervised and unsupervised.

Supervised methods mainly focus on the construction of a machine learning classifier to predict the reliability of parses or dependencies based on various kinds of features both on syntactic and semantic level. Yates et al. (2006) created WOODWARD which is a Web-based semantic filtering system. Kawahara and Uchimoto (2008) built a binary classifier that classifies each parse of a sentence as reliable or not. Among supervised methods, ensemble approaches were also proposed. Reichart and Rappoport (2007) detected parse quality by a Sample Ensemble Parse Assessment (SEPA) algorithm. Another similar approach proposed by Sagae and Tsujii (2007) also selected high quality parses by computing the level of agreement on different parser outputs. Iwatate (2012) applied a tournament model on Japanese dependency parsing and then selected reliable dependencies by using SVM output. The work most related to ours is the work of Yu et al. (2008). They proposed a framework that selects high quality parses in the first stage, and then selected high quality dependencies from the filtered parses. In comparison, we consider that even some low quality sentences possibly contain high quality dependencies. Also, we take into account other aspects that can affect high quality dependency classification and create a new set of linguistic features for classification.

Also, unsupervised algorithms for detecting reliable dependency parses were proposed. Reichart

International Joint Conference on Natural Language Processing, pages 947–951,
Nagoya, Japan, 14-18 October 2013.

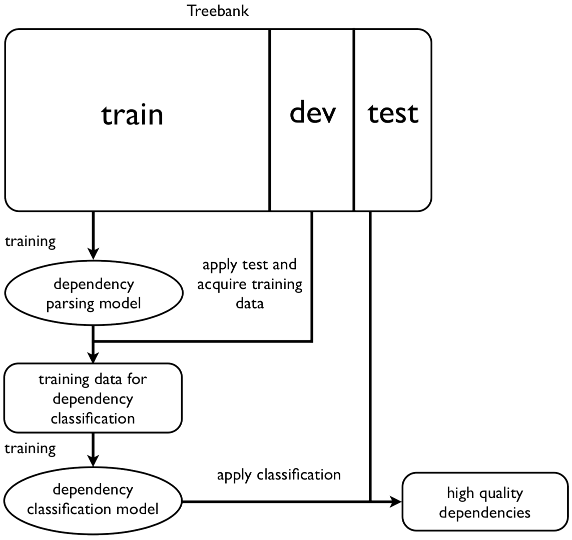
Treebank
train
dev
test
training
dependency parsing model
apply test and acquire training data
training data for dependency classification
training
dependency classification model
apply classification
high quality dependencies

Figure 1: Overview of high quality dependency selection

and Rappoport (2009) proposed an unsupervised method for high quality parse selection, which was based on the idea that syntactic structures that are frequently created by a parser are more likely to be correct. Dell'Orletta et al. (2011) proposed ULISSE (Unsupervised LInguiSticallydriven Selection of dEpendency parses), which uses an unsupervised method in a supervised parsing scenario. Although unsupervised methods may solve the domain adaption issue and do not use any annotated data, the accuracy of selected parses, which is under 95%, still needs to be improved for knowledge acquisition tasks.

3 High Quality Dependency Selection

In this section, we present a framework of highly reliable dependency selection from automatic parses. Figure 1 shows the overview of our approach. We use a part of a treebank to train a parser and another part to train a binary classifier which judges a dependency to be reliable or not.

3.1 Training Data for Dependency Classification

We collect training data from the same corpus which is also used in dependency parsing in the first stage. First, the training section is used to train a dependency parser and the development section is used to apply dependency parsing using the model trained by training section. From the parses of the development section, we acquire training data for dependency classification by labeling each dependency according to the gold standard data. All the correct dependencies are defined as reliable and vice versa.

3.2 Features for Dependency Classification

Most basic features consider that word pairs are much less likely to have a dependency relation when there are punctuation between them. On the other hand, based on the fact that dependencies with longer distance always show worse parsing performance (McDonald and Nivre, 2007), distance is another important factor that reflects the difficulty of judging whether two words have a dependency relation. Yu et al. (2008) used the features mentioned above and PoS features except the word features and did not use the context features, which are described later.

In addition to these basic features, we consider context features that are thought to affect the parsing performance. Table 2 lists these context features. In some more complex cases, it is also necessary to observe larger span of context. In order to learn such linguistic characteristics automatically, besides POS tags the head and modifier in a dependency, we also use their preceding and following one and two words along with their POS tags.

Another important fact is that verbal phrases in the dependency tree structure of a parse are normally the root node of the whole dependency tree or the parent node of a subtree. When a word pair that contains a verbal phrase between them, the two words are always on different sides of a parent node. Thus, these kinds of word pairs will always have no dependency link between them. This leads to the fact that argument pairs that have a verb between them rarely have a dependency relation. Observing whether there are verbal phrases between a head-modifier pairs can help judge whether the dependency between them is reliable.

The input of our high quality dependency selection method is a dependency tree. It is very natural to use tree-based features to identify the quality of dependencies. Based on a head-modifier dependency pair, we observe modifier's modifiers, i.e. children nodes. We use the leftmost and rightmost of children nodes to represent all the children nodes. We also take head's parent node into consideration, which we call a modifier's grandparent node. Furthermore, children nodes of the grandparent node which we call a modifier's uncle nodes are also considered as other features. Similarly, we use leftmost and rightmost uncle nodes.

4 Experiments

4.1 Experimental Settings

We first experiment on English, Chinese and Japanese. For English, we employ MSTparser[1] as a base dependency parser and use sections 02 to 21 from Wall Street Journal (WSJ) corpus in Penn Treebank (PTB) to train a dependency parsing model. Then, we use section 22 from WSJ to apply the dependency parsing model to acquire the training data for dependency classification. MXPOST[2] tagger is used for English automatic POS tagging. For Chinese, we use CNP (Chen et al., 2009) parser to train a dependency parser using section 1 to 270, 400 to 931 and 1001 to 1151 from Penn Chinese Treebank (CTB). Sections 301 to 325 are used to apply dependency parsing to acquire training data for dependency classification. We use MMA (Kruengkrai et al., 2009) to apply both segmentation and POS tagging. Different from the previous two languages which take *words* as the basic unit, experiments on Japanese are based on the unit of the phrase segments *bunsetsu*. We first use JUMAN[3] for Japanese morphological analysis. Then KNP[4] is utilized for Japanese dependency parsing. Section 950112, 950113 and 9509ED from Kyoto Corpus are used to apply dependency parsing and acquire training data for dependency selection.

We employ SVM-Light[5] with polynomial kernel (degree 3) to solve the binary classification. In order to compare with previous work by Yu et al. (2008), we use the basic feature set as a baseline. For English, section 23 from WSJ is used as a test set. Section 271 to 300 from CTB, and section 950114 to 950117 and 9510ED to 9512ED from Kyoto Corpus are used to test the classification approach in Chinese and Japanese, respectively. are used to test the classification approach in Chinese and Japanese respectively.

According to the output of the SVM, we only select dependencies that have the score higher than a threshold. Precision is calculated as the ratio of correct dependencies in retrieved ones. Recall is the ratio of correct dependencies in total. In Chinese and Japanese, we treat incorrect segmentations as incorrect dependencies. Note that the maximum recall value equals the precision of base dependency parser without dependency selection.

4.2 Experimental Results

4.2.1 Effectiveness of Dependency Selection

Figure 2 shows the precision-recall curves of the classification using SVM for three languages. In these graphs, 'basic' means the method using the basic features, 'context' stands for the mothod with context information, and 'context+tree' means the method with additional tree-based features.

One of the biggest problems that most data-driven parsers are facing is the domain adaption problem. When they are applied to a text of a different domain, their accuracy decreases significantly. We applied the dependency parsing model trained on WSJ to the Brown corpus, and obtained an unlabeled attachment score of 0.832, which is significantly lower than the in-domain score by 8.1%. We applied the same dependency selection model trained on WSJ to the Brown corpus. Figure 4 shows the precision-recall curves of dependency selection on the Brown corpus. From the results, we can see that when the recall is 40% for example, high quality dependencies with a precision of over 95% can be acquired. This shows that our method works well on data from different domains. This fact creates a good way to acquire knowledge from a large raw corpus in different domains (e.g., the Web).

4.2.2 Statistics of Selected Dependencies

In this section, in order to know what kind of dependencies are mainly selected, we show an investigation on the distribution of types of dependencies. Each dependency type in English and Chinese is represented by the coarse-grained POS pairs (the first two characters of POS names). Japanese dependencies are represented by the translated POS tags of *Bunsetsu* pairs. Figure 3 shows the statistics of POS pairs in three different languages. the leftmost graphs are drawn without selection. The middle and right graphs stand for

[1] http://sourceforge.net/projects/mstparser/

[2] http://www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html

[3] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[4] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

[5] http://svmlight.joachims.org

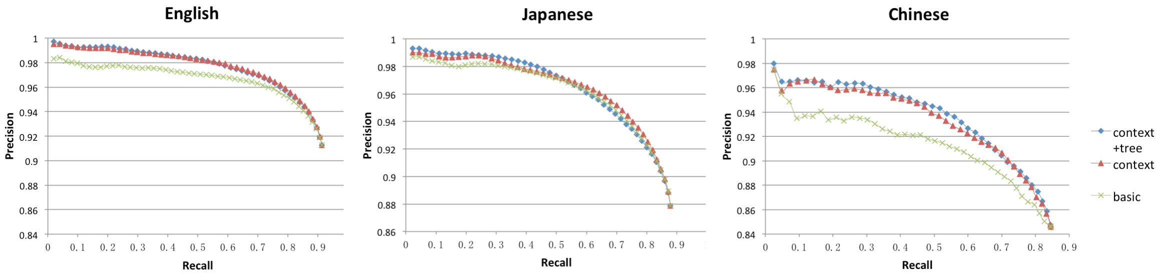

Figure 2: Precision-recall curves of dependency classification for English (left), Japanese (middle) and Chinese (right)

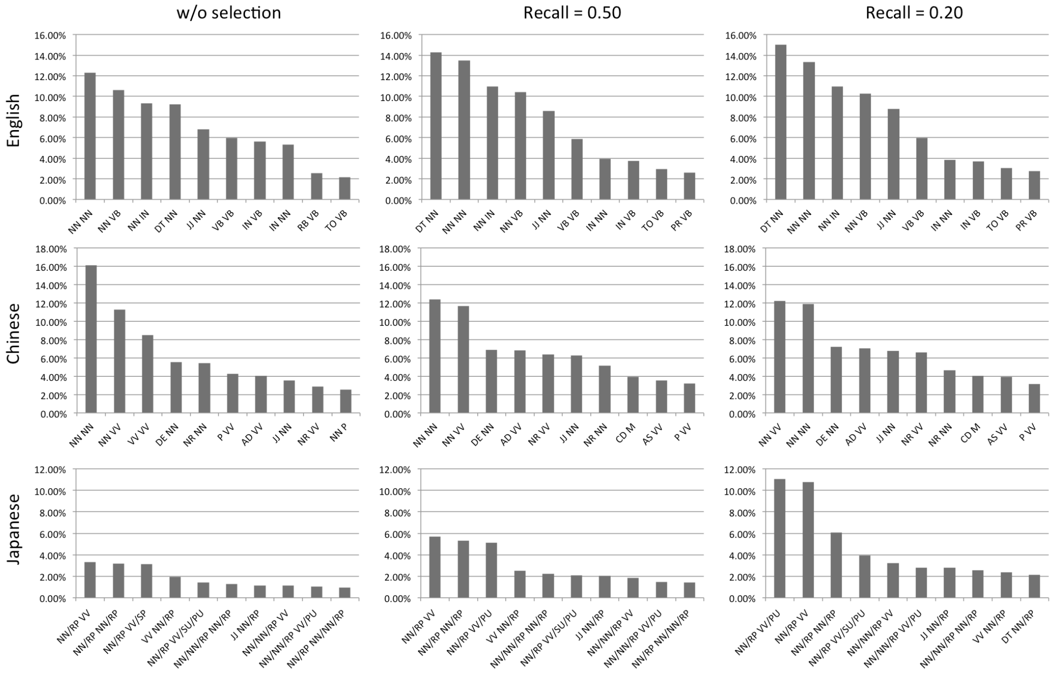

Figure 3: Statistics of POS tags of dependencies in different languages: dependencies without selection (left), dependencies when recall is 50% (middle), dependencies when recall is 20% (right)

the dependencies selected under different thresholds (i.e., recall is 20% and recall is 50% respectively). We found that dependencies with nouns are dominant in all the types for all the languages. Secondly, dependencies related to verbs which are very informative patterns account for a large proportion.

5 Conclusion and Future Work

In this paper, we proposed a classification approach for high quality dependency selection. We created new sets of features to select highly reliable dependencies from each parse through a parser. The experiments showed that our method worked for in-domain parses and also out-of-domain parses. We can extract high quality dependencies from a large corpus such as the Web and subsequently assist knowledge acquisition tasks, such as subcategorization frame acquisition and case frame compilation (Kawahara and Kurohashi, 2010), which depends highly on the parse quality. We also plan to use a bootstrapping strategy to improve a dependency parser based on acquired high quality knowledge from large corpora.

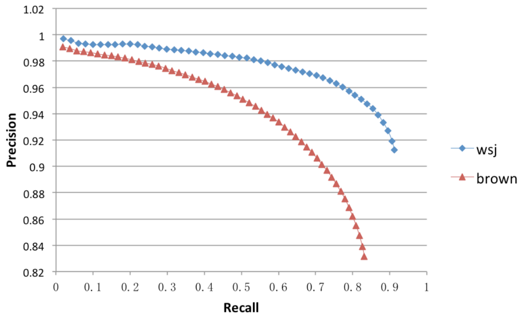

Figure 4: Precision-recall curves of dependency selection on Brown corpus

References

Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimota, and Kentaro Torisawa. 2009. Improving dependency parsing with subtrees from auto-parsed data. In *Proceedings of EMNLP 2009*, pages 570–579.

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2011. Ulisse: an unsupervised algorithm for detecting reliable dependency parses. In *Proceeding of CoNLL 2011*, pages 115–124.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi. 2011. Extracting paraphrases from definition sentences on the web. In *Proceedings of ACL 2011*, pages 1087–1097.

Masakazu Iwatate. 2012. *Development of Pairwise Comparison-based Japanese Dependency Parsers and Application to Corpus Annotation*. Ph.D. thesis, NAIST, Japan.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of HLT-NAACL 2006*, pages 176–183.

Daisuke Kawahara and Sadao Kurohashi. 2010. Acquiring reliable predicate-argument structures from raw corpora for case frame compilation. In *Proceedings of LREC 2010*, pages 1389–1393.

Daisuke Kawahara and Kiyokata Uchimoto. 2008. Learning reliability of parses for domain adaptation. In *Proceedings of IJCNLP 2008*, pages 709–714.

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An error-driven word-character hybrid model for joint chinese word segmentation and pos tagging. In *Proceedings of ACL-IJCNLP 2009*, pages 513–521.

Ryan McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of EMNLP-CoNLL 2007*, pages 122–131.

Roi Reichart and Ari Rappoport. 2007. An ensemble method for selection of high quality parses. In *Proceedings of ACL 2007*, pages 408–415.

Roi Reichart and Ari Rappoport. 2009. Automatic selection of high quality parses created by a fully unsupervised parser. In *Proceedings of CoNLL 2009*, pages 156–164.

Stijn De Saeger, Kentaro Torisawa, Masaaki Tsuchida, Jun'ichi Kazama, Chikara Hashimoto, Ichiro Yamada, Jong Hoon Oh, István Varga, and Yulan Yan. 2011. Relation acquisition using word classes and partial patterns. In *Proceedings of EMNLP 2011*, pages 825–835.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with lr models and parser ensemble. In *Proceedings of EMNLP-CoNLL 2007*, pages 408–415.

Alexander Yates, Stefan Schoenmackers, and Oren Etzioni. 2006. Detecting parser errors using web-based semantic filters. In *Proceedings of EMNLP 2006*, pages 27–34.

Kun Yu, Daisuke Kawahara, and Sadao Kurohashi. 2008. Cascaded classification for high quality head-modifier pair selection. In *Proceedings of NLP 2008*, pages 1–8.