A Framework for Compiling High Quality Knowledge Resources From Raw Corpora

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

The identification of various types of relations is a necessary step to allow computers to understand natural language text. In particular, the clarification of relations between predicates and their arguments is essential because predicate-argument structures convey most of the information in natural languages. To precisely capture these relations, wide-coverage knowledge resources are indispensable. Such knowledge resources can be derived from automatic parses of raw corpora, but unfortunately parsing still has not achieved a high enough performance for precise knowledge acquisition. We present a framework for compiling high quality knowledge resources from raw corpora. Our proposed framework selects high quality dependency relations from automatic parses and makes use of them for not only the calculation of fundamental distributional similarity but also the acquisition of knowledge such as case frames.

Keywords: Dependency selection, Knowledge acquisition, Case frames

1. Introduction

In natural language processing (NLP), rich knowledge is a strong backup for various kinds of tasks ranging from fundamental analysis, such as dependency parsing and word similarity calculation, to multilingual applications, such as machine translation. For instance, in the classic example of dependency parsing, "saw a girl with a telescope," there is an ambiguity problem of which argument the prepositional phrase, 'with a telescope,' is modifying. It would be much more easier to judge that the prepositional phrase is modifying the verb if knowledge of a case frame "someone sees someone/something with telescope/binocular," is available.

Manually constructing knowledge resources is very costly not only for construction but also for updating. Furthermore, manually constructed knowledge resources are always suffering from low coverage. As a result, automatic knowledge acquisition from large raw corpora has been actively studied recently. Knowledge is often acquired from syntactic analyses, such as constituency parses and dependency parses. In particular, dependency parsing has been used for many 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 precise knowledge. If all such dependency parses are used for knowledge acquisition, they produce a noisy knowledge resource, which leads to the deterioration of subsequent tasks using the knowledge base. Furthermore, if one tries to apply a method of knowledge acquisition to difficult-to-analyze languages like Chinese and Arabic, the quality of the resulting knowledge will get much worse.

During the dependency parsing process, a dependency parser tends to judge certain types of dependency relations with high accuracy. On the other hand, some specific types of dependency structures are relatively difficult for a parser to analyze correctly. As a result, a parser will produce automatic parses in different quality according to different properties of dependency. 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 framework for knowledge construction from high quality dependencies that are selected from automatic dependency parses. To our knowledge, there have been no studies that use high quality partial parses for knowledge acquisition. We experiment on English and Chinese using the same framework.

2. Related Work

To assist many kinds of text understanding task and other fundamental analysis, many language resources were built in previous studies. For example, there were manually constructed languages resources called FrameNet (Boas, 2002) and PropBank (Kingsbury and Palmer, 2003), which are corpora with verbal annotations. Even though both FrameNet and PropBank can provide annotated data as gold standard for many NLP applications, manually construction can hardly avoid the fact that the coverage of each knowledge repository is relatively low.

There have been many studies on automatic construction of such knowledge such as Subcategorization frames (Korhonen et al., 2006). subcategorization frames were proposed to represent the relations between the verbs and their syntactic arguments in the text. Subcategorization frames do not concern the meaning of each argument but focus on the argument patterns of the verb, and judge whether a certain kind of pattern makes sense on the frequency of this type pattern extracted from corpora. However, lack of detail information of the arguments is the biggest limitation of subcategorization frames and makes it less effective to use subcategorization frames to assist other applications in NLP. In later period, case frames (Kawahara and Kurohashi, 2006) for Japanese have also been automatically constructed. Although syntactic parsing plays a very important role in NLP,

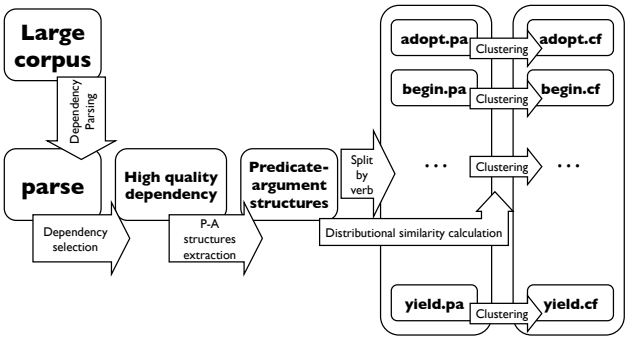

Figure 1: Overview of Case Frame Construction

there is still much information that cannot be efficiently indicated due to the characteristics of different language. In the case of Japanese, because of the special language characteristic such as omission of case components and case markers, case structure analysis is essential and construction of case frames became a very important issue. Different from subcategorization frames, case frames provide not only the information of predicate but also its arguments along with their relations in the text. Kawahara and Kurohashi (2006) first built a large-scale raw corpus from the Web and applied parsing. To avoid the bad effect of automatic parsing errors, they made use of Japanese-specific rules to extract reliable predicate-argument structures from the automatic parses. To address the problem of verb sense ambiguity, they finally applied a clustering process to acquire wide-coverage case frames with different usages for each verb. Their case frames consist of word examples but not semantic features. Instead of creating different language specific filtering rules, our proposed framework compiles knowledge bases from automatically selected dependenies from automatic parses.

3. Framework for Compiling High Quality Knowledge Resources

In this paper, we focus on the automatic compilation of two knowledge resources: a distributional thesaurus and case frames based on the distributional thesaurus. Figure 1 shows the overview of our framework. In this framework, dependency parsing is first applied to a large raw corpus. To overcome the issue of the imperfect performance of a dependency parser, we select high quality dependencies from the automatic parses. Then, we utilize the high quality dependencies to extract predicate-argument structures and construct a distributional thesaurus based on them. Finally, we cluster predicate-argument structures to produce case frames for each predicate.

The following subsections describe the details of these steps.

3.1. High Quality Dependency Selection

Instead of directly using all the automatic parses, we apply a dependency selection approach and then extract predicate-argument structures from the high quality dependencies. This idea is based on the fact that, a dependency parser tends to analyze different types of text in different level of performance. Take the two sentences "they eat salad with a fork" and "they eat salad with sauce" as examples. These examples have the PP-attachment ambiguity problem, which is one of the most difficult problems in parsing. The two prepositional phrases 'with a fork' and 'with sauce' depend on the verb 'eat' and the noun 'sauce,' respectively. However, these two cases can hardly be distinguished by a dependency parser due to the lack of knowledge like case frames. Therefore, we want to judge this kind of structure to be unreliable. Consider another similar sentence "they eat it with a fork." Since the prepositional phrase 'with a fork' cannot depend on the pronoun 'it' but only on the verb phrase 'eat,' this case can be clearly judged as a highly reliable dependency.

We employ the high quality dependency selection approach described in Jin et al. (2013), which shows good performance not only in in-domain cases but also out-of-domain case. This method first trains a base parser using a part of treebank. Then, they apply dependency parsing on the raw text of another part of the same treebank in order to collects training data for dependency selection according to the gold-standard annotations. They use context features and tree-based features, which are thought to affect the selection approach. Then, SVM is employed to solve the binary classification problem that classifies if each dependency is high quality or not. We do not apply a high quality parse selection approach (Yu et al., 2008) because we believe that there still exist many high quality dependencies even in low quality parses which could be also informative.

3.2. Predicate-argument Structure Extraction

Predicate-argument structures mainly capture the syntactic relations between a predicate and its arguments. Building wide-coverage case frames for each verb is basically to apply clustering predicates-argument structures of each predicate. Japanese predicate-argument structures have been successfully extracted and used for case frame construction (Kawahara and Kurohashi, 2006), where each argument is represented as its case marker in Japanese, such as 'ga', 'wo' and 'ni'. However, for other languages such as English and Chinese, there are no such case markers that can help clarify syntactic structures. Therefore, instead of using case markers like in Japanese, we represented each argument by its syntactic surface case (i.e., subject, object, prepositional phrase, etc.). Kawahara and Kurohashi (2010) used a chunking-based approach for large-scale predicate-argument structure acquisition. Instead of capturing dependency relations, this method uses language-specific filtering rules and only selects surrounding arguments and lacks multilinguality.

In order to extract high quality predicate-argument structures from all kinds of structure, we define a simple set of extraction rules for each language. First, to reduce the complexity of multi-verb cases, we focus on the last predicate in each sentence, which is chosen to be the predicate of this sentence. We only maintain the arguments which hold a dependency relation with the predicate. From the position of the predicate, the nearest preceding noun argument is selected as the subject. The following noun arguments are seen as the objects (direct object and indirect objects). A prepositional phrase is represented as a pair of preposition

and its argument (e.g., *pp:in:park*). Surface cases of other arguments are represent in their lower case of POS tags. We also distinguish the active and passive voices of a verb. In English for example, we further see whether there exists a verb 'be', which is the head of the chosen predicate. If so, the predicate would be the combination of 'be' and the verb's passive form (e.g., *be_shown*). Then, the nearest preceding noun argument which has a dependency relation with the verb 'be'.

Similarly, Chinese predicate-arguement structures are also represented by surface cases. However, Chinese passive voice is little more complex than English. Chinese passive voice is basically marked by character '被(Bei)'. According to annotation criteria in Chinese Treebank, '被' has two types of POS tag in different situations. The following two examples explain this phenomenon.

- 公司(company) 被(Bei) 政府(government) 列为(list as) 十强(top ten)
- 公司(company) 被(Bei) 列为(list as) 十强(top ten)

The POS tag of characater '被' in the first sentence is 'LB' which is the abbreviation for 'Long Bei'. This stands for the long distance between '被' and the verb '列为'. In contrast, '被' in the second sentence is marked as 'SB' which is the abbreviation for 'Short Bei'. In the case where '被' is directly adjacent to the verb, its POS tag is 'SB'. In other cases, the POS tag of '被' will be 'LB'. As in the first example, '政府', which is a modifier of verb '列为', is labeled as the subject. The argument '公司' which is the modifier of '被/LB' in the first example, and the modifier of the verb '列为' in the second example, becomes the direct object. There is another special case called '把(Ba)' in Chinese which indicates the direct object:

- 美国(America) 把(Ba) 此(This) 作为(take as) 窗口(window)

In this example, argument '此' is indicated as the direct object of the verb '作为', even though it appears before the verb. Argument '美国' which is a modifier of '把' became the subject of the verb '作为', even though they have no direct dependency relation.

3.3. Distributional Similarity

To measure the similarity between words, we use distributional similarity calculated from the same predicate-argument structures as described above. Distributional similarity is based on the hypothesis that words with similar semantic features always share the similar contexts (Hindle, 1990). The similarity between two objects is defined to be the amount of information contained in the commonality between the objects devided by the amount of information in the descriptions of the objects. We utilize a method that determines word similarity on the basis of a metric derived from the distribution of verb and its arguments (subject, object etc.) in a large text corpus (Lin, 1998), which is purely syntax-based similarity measurement. We acquire predicate-argument pairs and each of then is in the form of dependency triple with its frequency:

- freq(beer, subj-of, make) = 28
  freq(beer, subj-of, have) = 23
  ...
  freq(beer, obj-of, drink) = 20
  freq(beer, obj-of, make) = 10
  ...
- freq(wine, subj-of, make) = 30
  freq(beer, subj-of, spray) = 25
  ...
  freq(beer, obj-of, drink) = 16
  freq(beer, obj-of, make) = 10
  ...

Where the second element represents the syntactic relation between the first element and the third element. Since the calculation is based on predicate-argument structures from large corpora, the quality of predicate-argument structures will directly influence the quality of similarity calculation. Therefore, the selection of high quality dependencies is applied.

3.4. Case Frame Construction

Knowledge bases that mainly focus on predicates have been constructed (Korhonen et al., 2006). They mainly apply clustering to cluster semantically similar verbs (Reichart and Korhonen, 2013). However, in order to distinguish different semantic usages of each verb, internal clustering for each verb such as case frames is needed. For each verb, we apply a semantic clustering on its all predicate-argument structures. The clustering approach is similar to Japanese case frame construction (Kawahara and Kurohashi, 2006), which contains two steps.

First, we rank each surface case in a predicate-argument structure by pre-defined importance order and then choose the most important key argument which is seems to be most informative (e.g., direct object, subject, and prepositional phrase for English). Most of the time the direct object is seemed to be most important except for some predicate-argument structures with no direct object, because the meaning of an ambiguous verb is generated mostly by the co-composition of verb and object (Tsubaki et al., 2013). For each verb, all the predicate-argument structures that share the same key argument are clustered in the first stage to be the initial clusters. Secondly, we calculate the similarity between initial clusters by considering two aspects: 1. Alignment level; 2. Weighted case similarity. Alignment level is represented by the ratio of common surface cases two initial clusters are sharing. This actually indicates the syntactic similarity between two initial clusters. Weighted case similarity is a semantic level measurement which first considers word similarity of all the instances in the common surface case. Then word similarity is used to represent the case similarity between the common surface cases of both initial cluters. Without using any other additional thesauruses, word similarity we use for case frame compiling is actually calculated from previously compiled predicate-argument structures. Also considering that surface cases with more instances play more important roles, we define weighted case similarity according to the number of instances in each surface case.

During the clustering process, initial clusters are considered to be the smallest units. As new coming initial clusters will be gradually merged into different bigger clusters which

	En (in-domain)	En (out-of-domain)	Cn (in-domain)
Recall=20%	0.993	0.981	0.961
Recall=50%	0.983	0.951	0.945
Base parser	0.913	0.832	0.846

Table 1: Precision of selected dependencies under different criteria

contain numbers of initial clusters, we take the longest distance (i.e., the smallest similarity between a new coming initial clusters and all the grouped initial clusters in a big cluster) as the similarity between new coming initial cluster and the big cluster.

4. Experiments

4.1. Experimental Settings

For English, we employ MSTparser[1] as a base dependency parser and use sections 02 to 21 from the Wall Street Journal (WSJ) corpus in Penn Treebank (PTB) to train a dependency parsing model. We use section 22 from WSJ to acquire the training data for dependency selection. MXPOST[2] tagger is used for English POS tagging. Brown corpus is used to evaluate the out-of-domain performance of dependency selection.

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 5.0 (CTB). Sections 301 to 325 are used to acquire training data for dependency classification. We use MMA (Kruengkrai et al., 2009) to apply both segmentation and POS tagging. We employ SVM-Light[3] with polynomial kernel (degree 3) to solve the binary classification. From the output SVM score for each dependency, we only select the dependencies as high quality which have higher SVM scores than a threshold.

The distributional similarity is calculated from exactly the same predicate-argument structures we are using for case frames compilation. To show the effectiveness of high quality dependency selection approach in distributional similarity calculation, we calculate distributional similarities under three different sets of predicate-argument structures (i.e. without selection; a recall of 50%; a recall of 20%). For English, we employ Wordsim353[4] data set for evaluation. Wordsim353 is a gold-standard data set which has a human-assigned similarity between each word pair. For Chinese, we use a set of manually constructed gold-standard data[5] for Chinese word similarity evaluation, which contains more than 500 word pairs.

We use a large scale Web corpus which contain 200 million sentences for English case frame construction. For Chinese, we use five million sentences from the Chinese Gigaword.

4.2. Experimental Results

Table 1 shows the precisions for both of the languages in two different selection thresholds. From the results, we can see that it is possible to select higher quality dependencies (e.g., more than 98% for English) when we lower the recall. Also the out-of-domain case shows promising results, which mean that it is possible to apply high quality dependency selection process to raw text from different domains such as the Web. At the mean time, for it is relatively easy to acquire large scale of Web text, low recall can be compensated by the size of large corpus.

Figure 2 shows the spearman values of the distributional similarity calculations under the three criteria mentioned above. As we can see in the experimental result, performance of distributional similarity calculation can be improved by selecting high quality dependencies especially for English. However, the result for Chinese shows that, sometimes less selection has better performance on word similarity. We consider this phenomenon is due to deficiency in corpus size as the size of source corpus is much more smaller than English. Furthermore, although different types of dependencies can be select (not only 'DT NN' which is quite useless but also 'VV NN' which is more informative), It is still inevitable that some important information is lost during dependency selection process, especially while compiling from small size of corpus.

In case frame construction, we constructed 90 thousand types of predicates for English under the dependency selection threshold when the recall of selected dependencies is 20%, and 50 thousand types of predicates for Chinese under the dependency selection threshold when recall is 50%. Each English predicate contains around 60 case frames on average and each Chinese predicate contains around 50 case frames on average.

Table 2 and 3 give two examples of case frames in English and Chinese. In the English case frame for 'run', all the predicate-argument structures are clustered into different case frames to reflect different semantic usages. For example, the verb 'run' in case frame 'run(1)' basically means to 'execute' a programme, which is usually inanimate. The verb 'run' in case frame 'run(2)' means to 'take' a risk whose subject is often a person or an animate individual. In the Chinese example, verb '谢' is basically translated as 'thank' into English, which is actually represent by case frame '谢(2)'. Also '谢' can be the meaning of 'flower withering' or 'curtain call', whose ambiguity can be expressed in '谢(1)' and '谢(3)' correspondingly.

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

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

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

[4] http://alfonseca.org/eng/research/wordsim353.html

[5] http://www.cs.york.ac.uk/semeval-2012/task4/

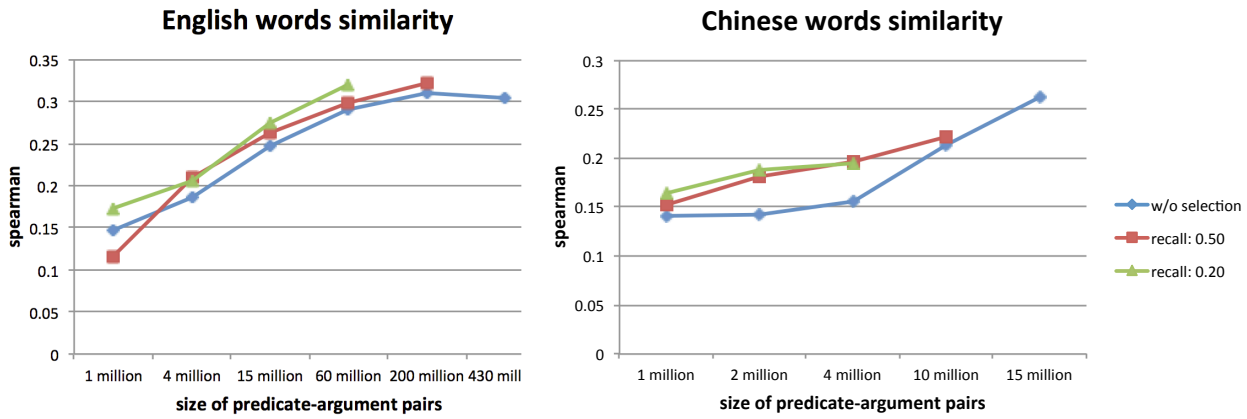

Figure 2: Spearman under different selection thresholds

verb	surface case	instance with frequency in original corpus
run(1)	sbj	programme:552, it:22
	pp:in	background:839, window:3, way:2, mode:2 ...
	rb	continually:16, smoothly:6, well:6 ...
	...	...
run(2)	sbj	he:10, it:3, they:3, i:2, individual:1
	obj	risk:336
	rb	also:2, ago:1, new:1
	...	...
...		

Table 2: Examples of English Case frames

verb	surface case	instance with frequency in original corpus
谢(1)	sbj	花儿(flower):14, 花(flower):22
	ad	都(all):16, 也(also):6
谢(2)	sbj	你们(you):1
	obj	您(you):8, 我(me):6
	ad	怎么(how):8, 多(very):1
谢(3)	sbj	大战(battle):1
	obj	幕(curtain):6
	ad	圆满(seccessfully):2, 也(also):1, 正式(officially):1
...		

Table 3: Examples of Chinese Case frames

5. Conclusion and Future Work

In this paper, we proposed a framework for automatic knowledge resource construction from high quality dependencies. The experiments showed that our dependency selection 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, which depend highly on the quality of automatic parses. We plan to enlarge the source corpora from different domains for larger scale case frame acquisition. Also, the balance between high quality dependency selection and important information maintenance is another important issue we need to work on in the future. In Chinese predicate-arguement structure construction, due to the language order of Chinese if relatively free, simple transformation rules sometimes can hardly precisely capture the surface cases. We want to make use the acquired knowledge to apply a self-correction process. We also plan to use a bootstrapping strategy to improve fundamental analysis such as dependency parsing itself based on acquired high quality knowledge from large corpora.

6. References

Boas, Hans C. (2002). Bilingual FrameNet dictionaries for machine translation. In Rodríguez, M. González and Araujo, C. Paz Suárez, editors, *Proceedings of the Third*

International Conference on Language Resources and Evaluation, volume IV, pages 1364–1371, Las Palmas.

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

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

Hindle, Don. (1990). Noun classification from predicate-argument structures. In *Proceedings of ACL 1990*, pages 268–275.

Jin, Gongye, Kawahara, Daisuke, and Kurohashi, Sadao. (2013). High quality dependency selection from automatic parses. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 947–951.

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

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

Kingsbury, Paul and Palmer, Martha. (2003). Propbank: The next level of treebank. In *Proceedings of Workshop Treebanks and Lexical Theories*.

Korhonen, Anna, Krymolowski, Yuval, and Briscoe, Ted. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th international conference on Language Resources and Evaluation*, pages 345–352.

Kruengkrai, Canasai, Uchimoto, Kiyotaka, Kazama, Jun'ichi, Wang, Yiou, Torisawa, Kentaro, and Isahara, Hitoshi. (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.

Lin, Dekang. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 768–774.

Reichart, Roi and Korhonen, Anna. (2013). Improved lexical acquisition through dpp-based verb clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 862–872.

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

Tsubaki, Masashi, Duh, Kevin, Shimbo, Masashi, and Matsumoto, Yuji. (2013). Modeling and learning semantic co-compositionality through prototype projections and neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 130–140.

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