Sentence Rephrasing for Parsing Sentences with OOV Words

Hen-Hsen Huang*, Huan-Yuan Chen*, Chang-Sheng Yu*, Hsin-Hsi Chen*, Po-Ching Lee†, Chun-Hsun Chen†

*Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan

†Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan, R.O.C.

E-mail: {hhhuang, csyu}@nlg.csie.ntu.edu.tw, {b00902057, hhchen}@ntu.edu.tw,
{albertlee, jeffzpo}@cht.com.tw

Abstract

This paper addresses the problems of out-of-vocabulary (OOV) words, named entities in particular, in dependency parsing. The OOV words, whose word forms are unknown to the learning-based parser, in a sentence may decrease the parsing performance. To deal with this problem, we propose a sentence rephrasing approach to replace each OOV word in a sentence with a popular word of the same named entity type in the training set, so that the knowledge of the word forms can be used for parsing. The highest-frequency-based rephrasing strategy and the information-retrieval-based rephrasing strategy are explored to select the word to replace, and the Chinese Treebank 6.0 (CTB6) corpus is adopted to evaluate the feasibility of the proposed sentence rephrasing strategies. Experimental results show that rephrasing some specific types of OOV words such as Corporation, Organization, and Competition increases the parsing performances. This methodology can be applied to domain adaptation to deal with OOV problems.

Keywords: Sentence Rephrasing; Named Entity; Dependency Parsing

1. Introduction

Out of vocabulary (OOV) is a "relative" concept. Those words not appearing in a lexicon are OOV words relative to this lexicon. Similarly, those words not appearing in a training corpus are also OOV to the corpus. In domain adaptation, words may be known in target domain, but are OOV in source domain. In real world, words are developed continuously, web data and social media data, in particular (Khan, et al., 2013; Øvrelid & Skjærholt, 2012). Furthermore, a training corpus is not always available for every domain, so that applications of models trained in a domain to process the text in a new domain are often needed. How to deal with such kinds of OOV words is an important research issue at various levels of natural language processing.

The information from an OOV itself such as morphemes and the contextual information around the OOV are both useful to predict its lexical/syntactic properties. Previous researches have employed such clues to deal with Chinese segmentation, named entity (NE) recognition, and part of speech (POS) tagging (Lin & Chen, 2006; Chan & Chong, 2013). POS, which demonstrates common properties of a group of similar items, is important for parsing sentences with OOV words. Dependency parsers often employ various features such as word form, POS, dependency type, and so on to predict the structure of a sentence (Nivre & Kubler, 2006; Wang & Zhang, 2010). If words in test sentences are unknown relative to a training corpus, their word forms are less useful in the structure prediction.

Sentence rephrasing aims to form a new sentence to convey a similar meaning as the original one, but has the better representation for some specific goals, e.g., easy understanding, better performance, and so on. Having certain words replaced with others is one of the common rephrasing operations. In Chen et al's simplification-translation-restoration framework (2012), they identify the domain-specific segments and simplify them into more general expressions for statistical machine translation (SMT). In general, simplification is regarded as some kind of rephrasing. It deals with the problem of SMT systems which cannot gather the statistical evidence of a segment containing OOV words.

This paper follows the general view of rephrasing methodology. We aim to replace an OOV word in a sentence with some word of the same NE type in the training set to have better parsing performance for this sentence. For example, the word 保通社 (bǎo tōng shè), the abbreviation of the news agency 保加利亞通訊社 (Bulgarian News Agency), is an OOV for the dependency parser. For the test instance containing this word as (S1), our rephrasing algorithm replaces it with the frequent word 新華社 (Xinhua News Agency), which is the name of a news agency.

(S1) 保加利亞外長米哈伊洛娃 27 日在接受保通社採訪時表示，1999 年保加利亞外交重點之一，是積極謀求地區穩定，解決科索沃危機。(In an interview with the Bulgarian news agency On 27, Bulgarian Foreign Minister Mikhailova said that one of the Bulgarian diplomatic focus in 1999 is actively seeking to regional stability and to resolve the Kosovo crisis.)

Next, the dependency parser will parse the rephrased sentence. In the end, the original word 保通社 (bǎo tōng shè) will be restored in the parsing result.

Chen & Bai (1998) examined 3 million words in Sinica corpus[1], and found that abbreviations, proper names, derived words, compounds, and numeric type compounds are 5 major types of OOV words in Chinese. For each type of OOV words, they are usually composed of specific morphemes. Such morphemes provide important clues to predict the type of an OOV. To demonstrate the feasibility of the proposed rephrasing method, named entities (NEs), which are very common OOV words (Chen & Bai, 1998), are regarded as our experimental targets.

2. Dependency Parser

Chinese Treebank 6.0 [2] (CTB) is adopted in the experiments. Proper nouns (NRs) denote NEs like persons, locations and organizations. We count the frequency of words with NR tag in CTB and partition it into training set and test set based on the frequency of NEs. There are 4,622 and 3,719 NEs occurring once and more than once in CTB, respectively. They form SING (single-occurrence) and MULT (multiple-occurrence) sets. Total 3,279 sentences contain at least one NE in the SING set. In addition, 350 of these 3,279 sentences also contain at least one NE in the MULT set. For ensuring the NEs in the test set are OOV relative to the training set, these 350 sentences are removed from the test set. In the experimental setup, we would like to observe the behaviors of parsing sentences with the OOV words.

We employ the CNP package[3] (Chen et al., 2009) to train a Chinese dependency parser with the above training set, and adopt the MaltEval tool[4] (Nilsson and Nivre, 2008) to evaluate the Chinese dependency parser. The 2,929 sentences containing at least one OOV NE are fed into the parser. Here we assume the correct POS is assigned to every word in a sentence. That is, all the OOV NEs will have the tag NR (proper noun). Because these NEs are OOV in the training set, their word forms cannot be used as features in the training process. We aim to find their effects on the parsing performance.

3. Rephrasing Strategies

Table 1 shows parts of the NE types (Sekine, 2008) adopted in this paper. In this scheme, there are four types including person, location, organization and others, and some subtypes for each type, e.g., persons with Chinese names and foreign names. The 2nd column lists some Chinese examples for each NE type. The 3rd and the 4th columns show the distributions of each NE type in the test and the training sets, respectively. For example, 40.07% of NEs in the test set are Chinese names, while Chinese names occupy 14.92% of NEs in the training set. Approaches of employing morphemes and contextual cues to recognize the type of an OOV NE have been proposed (Lin & Chen, 2006). The following presents three strategies to deal with OOV NEs.

(1) The word form of an OOV NE is not changed, and the corresponding POS is set to NR (proper noun). This strategy is regarded as a baseline.

(2) The word form of an OOV NE is rephrased as the highest frequent word of the same type as the OOV NE in the training set, and the corresponding POS is set to NR.

(3) The word form of an OOV NE is rephrased as a word selected from the training set by information retrieval (IR) approach, and the corresponding POS is set to NR. IR model is adopted to find the word of the most similar context with the test sentence to rephrase the OOV NE. Each sentence in the training set is indexed by the words in the sentence. The words in a test sentence are used to query the training set. Sentences are ranked by the TF-IDF function of the adopted IR model. The similarity scores are summed over the same NE instances in the training set. The NE of the highest score will be selected to rephrase the OOV NE in the test sentence.

4. Results and Discussion

Table 2 shows the performance of the dependency parser when POS is given. Each token is regarded as correct if both dependency label and head are correctly determined. Besides the above labeled attachment scores (LAS), the percentage of words that have the correct head, i.e., unlabeled attachment scores (UAS), is also computed. Token accuracy in LAS (UAS) for an NE type is defined as correct tokens in LAS (UAS) divided by number of tokens of that type.

The 2nd column shows the number of test instances for each NE type. Chinese name, foreign name, city (county), corporation, other location, and other organization are the top 6 NE types. The strategy of using the original OOV words, i.e., without rephrasing, is regarded as baseline. The 3rd column lists the LAS and the UAS for each type using Strategy 1. The 4th-6th columns demonstrate the LAS and UAS using Strategies (2) and (3), respectively. The accuracy in boldface means the rephrasing strategy has better performance than the baseline. The accuracy underlined means no change. The highest-frequency-based method (i.e., Strategy (2)) increases accuracy in the following NE types: state, providence, city, county, river, lake, sea, corporation, competition, artifact, and other organization. The default IR model of Lucent search engine is used for Strategy (3). There are two alternatives in the experiments. The difference between IR+V and IR+All POS is: the former employs verbs only for context similarity computation,

[1]http://rocling.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm
[2]http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2007T36
[3]http://alaginrc.nict.go.jp/cnp/index.html
[4]http://www.maltparser.org/malteval.html

Type			
Subtype	Chinese Examples	Test	Training
Person			
Chinese Name	吳京、趙明路、林青霞、林貝聿嘉、歐陽中石、譚某、老賈、姚老闆	40.07%	14.92%
Foreign Name	下山宏、巴沙爾・阿薩德、巴沙爾阿薩德、凱西米爾・奧耶・姆巴、Marcus	17.12%	8.76%
Location			
Country	巴哈馬、加國、大唐、冰島、德意志、荷、荷蘭	2.08%	42.09%
State, Province	華盛頓州、阿拉斯加、麻省、寬多－庫邦戈省、廣西省、贛省	1.13%	2.94%
City, County	東京都、大阪府、大安區、九江市、馬丁縣、卡倫加鎮、三芝鄉、名古屋	11.42%	13.72%
Address	八德路、三民街、花園道、慈雲巷	1.98%	0.12%
River, Lake, Sea	二仁溪、馬牧河、沅江、赤水、太陽湖、積水潭、北極海、麗都灣	1.50%	1.10%
Island	海南島、維爾京群島、馬來半島、威科斯島、威科斯	0.71%	0.23%
Mountain	大屯山、嘎夏布魯慕峰	0.69%	0.20%
Bridge	二水橋	0.08%	0.01%
Harbor	台中港、安特衛普港	0.69%	0.26%
Desert	巴丹吉林沙漠、戈壁、撒哈拉	0.05%	0.00%
Location Other	京滬、武漢關、昭忠祠、梅園、羅星塔、葛洲壩、赫倫堡、紅毛城	5.09%	4.72%
Organization			
Corporation	IBM、萬寶路、小美、飛利浦、中鋼	8.65%	2.62%
School	史丹佛、浙大、輔仁	0.76%	0.05%
Religion	義民廟、西來寺、慈天宮、媽閣、一貫道、玉皇大帝、閻王爺	0.79%	0.11%
News、TV	中新社、太陽報、新新聞、上海台	1.29%	1.96%
Competition	世運、亞洲杯、金球獎、英超聯賽	0.42%	0.28%
Team	大陸隊、弗魯米嫩塞隊、河南隊	1.03%	0.10%
Organization Other	中經院、天地會、民主進步黨、竹蓮幫、國安會	2.53%	5.27%
Others			
Music, Novel	大力水手、英烈千秋、鹿鼎記	0.45%	0.06%
Artifact	貝格爾艦、沙利文號	0.98%	0.32%
Event	二戰、五胡亂華、文藝復興	0.47%	0.16%

Table 1: Named Entity Ontology.

and the latter considers all words. Strategy 3 is workable for some specific NE types, but the performance decreases compared with the highest frequency based method. Table 2 shows that LAS of News and TV entity type using the highest-frequency-based strategy is worse than that using Strategies 1 and 3. In the experiments, the word of the highest frequency for this entity type is 新華社 (Xinhua News Agency), which is the official press agency of China. We further examine its tagging in CTB 6.0. Surprisingly, 378 occurrences of this NE were tagged with NR and 113 occurrences were tagged with NN in the training set. If we replace the OOV of News and TV type with the 2nd highest frequent word, i.e., 中央台 (CCTV), LAS increases from 0.653 to 0.714. In the training set, 中央台 (CCTV) was tagged NR 168 times and NN 4 times. This error analysis depicts that the quality of the training corpus affects the performance. Table 2 also shows that Strategies (2) and (3) do not outperform Strategy (1) on person type. In the highest-frequency-based strategy, 江澤民 (Jiang Zemin) and 克林頓 (Bill Clinton) are used to rephrase OOV words of Chinese and foreign names, respectively. They are former presidents of China and USA, respectively, thus have the highest mentions in the training set. Actually, OOV words of person type occupy 57.19% of test instances, but the person-typed NEs only occupy 23.68% in the training set. Intuitively, persons may play different roles and are more sensitive to the context in the descriptions. Thus, rephrasing strategies are still challenging for person types.

5. Conclusion

This paper proposes rephrasing strategies to deal with Chinese dependency parsing with OOV words. Named entities are taken as examples in the experiments. Experimental results show that rephrasing the OOV words in some specific NE types such as Corporation, Organization, and Competition improves the parsing performance. Compared to the IR-based strategies, the simple highest-frequency-based rephrasing strategy increases the parsing performance for more NE types. Restrictive rephrasing increases the overall parsing performance. This methodology can be applied to

Strategy →	#Test	Original OOV		Highest Frequency		IR+V		IR+All POS	
Subtype ↓	Instances	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
Chinese Name	1,519	0.821	0.883	0.821	0.880	0.811	0.873	0.819	0.881
Foreign Name	649	0.812	0.846	0.797	0.838	0.804	0.838	0.804	0.835
Country	79	0.785	0.785	0.772	0.772	0.772	0.772	0.785	0.785
State, Province	43	0.744	0.767	**0.767**	**0.791**	**0.767**	**0.791**	0.744	0.767
City, County	433	0.838	0.850	0.838	0.855	0.831	0.843	**0.843**	**0.852**
Address	75	0.787	0.800	0.787	0.800	0.773	0.787	0.773	0.787
River, Lake, Sea	57	0.737	0.754	**0.754**	**0.772**	**0.754**	0.754	**0.754**	**0.772**
Island	27	0.852	0.963	0.815	0.926	0.815	0.926	0.852	0.963
Mountain	26	0.769	0.808	0.769	0.808	0.769	0.808	0.769	0.808
Bridge	3	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667
Harbor	26	0.769	0.846	0.769	0.846	0.769	0.846	0.769	0.846
Desert	2	1	1	1	1	1	1	1	1
Location Other	193	0.886	0.886	0.876	0.881	0.876	0.876	0.870	0.876
Corporation	328	0.817	0.838	**0.829**	**0.851**	**0.820**	**0.841**	**0.823**	**0.845**
School	29	0.931	0.931	0.931	0.931	0.931	0.931	0.917	0.917
Religion	30	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800
News, TV	49	0.796	0.878	0.653	0.816	0.796	0.898	0.796	**0.898**
Competition	16	0.625	0.625	**0.750**	**0.750**	**0.688**	**0.688**	**0.688**	**0.688**
Team	39	0.897	0.923	0.897	0.923	0.846	0.872	0.872	0.897
Organization Other	96	0.854	0.885	**0.875**	**0.906**	0.802	0.833	0.823	0.854
Music, Novel	17	0.588	0.706	0.588	0.706	0.529	0.647	0.529	0.647
Artifact	37	0.730	0.811	**0.757**	**0.838**	0.703	0.784	0.676	0.757
Event	18	0.944	0.944	0.889	0.889	0.944	0.944	0.944	0.944

Table 2: Performance of Each OOV NE Type

domain adaptation to deal with OOV problems. In this paper, rephrasing is regarded as a preprocessing step before parsing. In the future work, we will integrate the features of NE types into parsers directly and introduce finer features such as properties of persons from knowledge graph in parsing.

6. Acknowledgements

This research was partially supported by National Science Council, Taiwan under NSC101-2221-E-002-195-MY3.

7. References

Chan, S.W.K. & Chong, M.W.C. (2013). Recursive Part-of-Speech Tagging Using Word Structures. In *Proceedings of Text, Speech, and Dialogue*. LNCS 8082, pp. 419-425.

Chen, H.B., Huang, H.H., Chen, H.H. & Tan, C.T. (2012). A Simplification-Translation-Restoration Framework for Cross-Domain SMT Applications. In *Proceedings of the 24th International Conference on Computational Linguistics* (COLING2012). Mumbai, India, pp. 545-560.

Chen, K.J. & Bai, M.H. (1998). Unknown Word Detection for Chinese by a Corpus-Based Learning Method. *Computational Linguistics and Chinese Language Processing*, 3(1), 27-44.

Chen, W., Kazama, J., Uchimoto, K. & Torisawa, K. (2009). Improving Dependency Parsing with Subtrees from Auto-Parsed Data. In *Proceedings of Conference on Empirical Methods in Natural Language Processing* (EMNLP2009). Singapore, pp. 570-579.

Khan, M., Dickinson, M. & Kuebler, S. (2013). Does Size Matter? Text and Grammar Revision for Parsing Social Media Data. In *Proceedings of with NAACL-HLT 2013 Workshop on Language Analysis on Social Media*. pp. 1-10.

Lin, M.S. & Chen, H.H. (2006). Constructing a Named Entity Ontology from Web Corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC2006). pp. 1450-1453.

Nilsson, J. & Nivre, J. (2008). MaltEval: An Evaluation and Visualization Tool for Dependency Parsing. In *Proceedings of International Conference on Language Resources and Evaluation* (LREC2008). pp. 161-166.

Nivrel, J. & Kubler, S. (2006). Dependency Parsing. Tutorial at COLING-ACL2006.

Øvrelid, L. & Skjærholt, A. (2012). Lexical Categories for Improved Parsing of Web Data. In *Proceedings of the 24th International Conference on Computational Linguistics* (COLING2012). pp. 903–912.

Sekine, S. (2008). Extended Named Entity Ontology with Attribute Information. In *Proceedings of International Conference on Language Resources and Evaluation* (LREC2008). pp. 52-57.

Wang, Q.I. & Zhang, Y. (2010). Recent Advances in Dependency Parsing. Tutorial at NAACL2010.