One Sense per Context Cluster: Improving Word Sense Disambiguation Using Web-Scale Phrase Clustering

Heng Ji
Computer Science Department
Queens College and Graduate Center
City University of New York
New York, NY 1167, USA
hengji@cs.qc.cuny.edu

***Abstract*—The performance of word sense disambiguation task is still limited by lexical context matching due to data sparse problem. In this paper we present a simple but effective method that incorporates web-scale phrase clustering results for context matching. This method is able to capture some semantic relations that are not in WordNet. Without using any additional labeled data this new approach obtained 2.11%-6.92% higher accuracy over a typical supervised classifier.**

Keywords-Word Sense Disambiguation, Clustering, Web-scale N-grams

I. INTRODUCTION

The problem of word sense disambiguation (WSD) has been extensively studied and applied to other natural language processing tasks such as machine translation (e.g. [1]; [2]) and text classification (e.g. [3]). Many methods used the feature sets based on context word matching, and so the knowledge acquisition bottleneck [4] still remains due to sparse data. In other words, the training data, especially the neighboring contexts for similarity matching, for each test instance may not be available. In order to learn a more robust WSD classifier, the feature sets for context matching should go beyond lexical level, and exploit syntactic and semantic information. Some recent literature (e.g. [5]; [6]; [7]) have been alleviating this bottleneck by using the semantic relations in WordNet [8]. For example, one can match the contexts between a training instance and a test instance by their synonym sets in WordNet. This approach has two main limitations: 1. It cannot address broader semantic relatedness; 2. It cannot address the semantic relations between two words with different part-of-speech tags.

In this paper we propose a more relaxed context matching approach. We apply the unsupervised phrase clustering method described in ([9], [10]) on web-scale n-grams to generate clusters. Then for each context word of a target word, we search for its closest cluster. This allows us to achieve more accurate context matching on the semantic cluster level instead of lexical level. Our experiment results show that this method can achieve significant improvements over a typical supervised WSD classifier. The rest of this paper is structured as follows. Section II presents the overall system architecture. Section III then describes the characteristics of the results from web-scale phrase clustering compared to WordNet. Section IV presents the experimental results. Section V compares our approach with related work and Section VI then concludes the paper and sketches our future work.

II. ALGORITHM OVERVIEW

[11] states that words that occurred in the same contexts tend to be similar. Following this intuition, most supervised WSD systems used context words surrounding the target word and the WordNet relatives (such as synonyms, hypernyms and hyponyms) of the target word as features. We propose to extend this hypothesis to *related* contexts – *one sense per context cluster*. In other words, two target words surrounded by related contexts tend to have the same sense. Furthermore, we hypothesize that the relatedness measure for contexts should be broader than those defined in WordNet. We shall apply a web-scale phrase clustering method to verify these hypotheses. We start with presenting the overall procedure of our approach as follows.

In the offline procedure, a distributed version of K-means clustering method ([9]; [10]) is applied to cluster the Google n-gram (n=5) corpus Version II, which can be viewed as a compressed summary of the web. Google n-gram Version II includes 207 billion tokens selected from the LDC-released Version I, consisted of 1.2 billion 5-grams extracted from about 9.7 billion sentences. All these 5-grams are automatically annotated with part-of-speech (POS) tags based on their original sentences.

Then we utilize the resulting clusters for WSD in a Naïve Bayesian classifier. For the trigram context words $\{w_{-3}, w_{-2}, w_{-1}, w_1, w_2, w_3\}$ surrounding each target word *t*, we search for their corresponding closest clusters generated from above: $\{c_{-3}, c_{-2}, c_{-1}, c_1, c_2, c_3\}$. If a word is not part of any clusters, we consider it as an independent (1-word) cluster. In this way, we can estimate more reliable probabilities for context words.

In the test procedure, we can determine the sense of a test target word *t* from its possible sense set *S* by computing the joint probabilities based on the cluster features:

$$sense(t) = \arg\max_{s \in S}(\prod_{i=-3}^{-1} p(c_i \mid s) \times \prod_{i=1}^{3} p(c_i \mid s) \times p(s))$$

where $p(c_i|s)$ is the probability of the cluster c_i appearing in the contexts of the training instances that indicate the sense *s*; and $p(s)$ is the probability of sense *s* occurred in the training data.

III. One Sense per Context Cluster

Now we will present some concrete examples of context words that support our hypothesis of "one sense per context cluster". These context words cannot be connected by any semantic relations defined in WordNet, however, they were successfully clustered by the web-scale phrase clustering method. We categorize them into the following types.

A. Cross-tag Synonyms

A notable advantage of web-scale clustering is that it can capture the relations among those words with different part-of-speech tags. For example, in the following sentences:

> *The two men arrived at Ajaccio on the Portugal, a battered old steamer which **moored** alongside a <head>palm</head> -lined quay*
>
> *The woodland, marching up the hill, vanished before it but reached an arm around to the west, fringing the road, and then ran behind it to the north, forming a long **backdrop** to the <head>palm</head> house and the terraces. Only the clock tower on the stables showed from behind the trees.*

Although the context words "*moored*" and "*backdrop*" have different part-of-speech tags, they are both clustered into the same set and thus indicate the same sense "*tree*" for the target word "*palm*".

B. Antonyms

The antonymy relation is usually not exploited in WSD. However, if two context words are antonyms and belong to the same semantic cluster, they tend to represent the alternative attributes for the target word. For example,

> *The cameras clicked, the reporters reached for their notebooks, the Principal and Lord James Douglas-Hamilton **ascended** in a <head>crane</head> to unveil a sign…*
>
> *I get on well with old Bert. A huge steel <head>crane</head> hook suddenly **descended** quietly between their faces and made them both leap back in alarm. Yanto shook his fist at the face of the crane driver grinning down at them through his cab window.*

"ascended" and "descended" are antonyms but they both reflect the main function of a "crane" with "machine" sense, and can be exploited to distinguish it from the other sense "bird".

C. Subsequent events

In addition, some context words often appear in a sequence of events and so can be clustered together.

For example, in the following sentences, "*drive→hanging→lifting*" are clustered into the same set by the web-scale clustering method because they all represent the event series normally acted by a "*crane*" with a "*machine*" sense:

> *He didn't want to know anything he wanted to **drive** the <head>crane</head> so he give me a start.*
>
> *However, the America's Cup is not about launches and, as even Gardini could see, the boat **hanging** from the <head>crane</head> could take the Spanish far.*

*The Customer shall provide such temporary roadways, footways, scaffolding and other equipment as may be necessary for the safe installation of the goods who has to bear the cost of equipment hire if a fork lift truck, or heavy **lifting** <head>crane</head>, is needed for the installation?*

Such subsequent events can also be acted by different agents and recipients. For example, "*sang*" and "*heard*" are clustered and both indicate the "*music*" sense of "*bass*":

*By profession I was an opera and oratorio singer, and I **sang** the <head>bass</head> solos in the Messiah in the Ulster Hall.*

*Then I **heard** Quigley's low <head>bass</head>, but couldn't make out what he was saying.*

D. Topically-Related Words

The web-scale clustering method can also generate many topically-related clusters for context matching. Table 1 presents some examples. For example, if "guitar" mostly indicates "music" sense of "bass" in the training data, once "playing" appears in the context of any test instance we can determine that "bass" is likely to have "music" sense.

TABLE I. EXAMPLES FOR TOPICALLY-RELATED CONTEXT WORD CLUSTERS

Target Word	Sense	Context Word Clusters
bass	music	{guitar, playing, hear, beat, voices,…}
palm	tree	{sit, lap, stood, waited, …}
motion	physical	{thrust, swinging, bumping, swaying, tucking, …} {passed, reached, standing, stopped, greeted, …}
tank	container	{revived, digging, escaping, …}

IV. EXPERIMENTAL RESULTS

In this section we will present the results of applying this web-scale clustering method to improve word sense disambiguation.

A. Data

We evaluated our approach on the TWA Sense Tagged Dataset [11] based on five-folder cross-validation. The corpus includes binary sense tagged sentences for six words (104 instances for "bass"; 201 instances for "motion"; 95 instances for "crane"; 188 instances for "plant"; 201 instances for "palm" and 201 instances for "tank").

B. Results and Discussions

For comparison we built a baseline system that uses the synonym sets in WordNet for cluster searching. Table 2 shows the overall performance of WSD results using two clustering methods. We can see that for all of the six words, our approach achieved significant improvement over the baseline.

TABLE II. OVERALL PERFORMANCE

Target Word	WordNet based Clustering	Web-scale Phrase Clustering
Bass	92.73%	95.61%
Crane	88%	90.11%
Motion	79.21%	84.19%
Palm	84.16%	88.14%
Plant	72.63%	79.55%
Tank	71.29%	73.78%

In addition, we conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on five folders for each target word. The results show that all the improvement using web-scale clustering is significant at a confidence level higher than 96.7%. It is also worth noting that the most significant improvement was achieved for the target word "plant" because no overlapped context clusters were generated for its two senses "factory" and "living".

We have also analyzed the cases for which the web-scale clustering method performed worse. The errors reveal both the shortcomings of automatic phrase clustering and consistent difficulties of using contextual similarity. Some common context clusters can indicate both senses of a target word, and thus cannot provide additional improvement. For example, in "…other than a **series** of anti-<head>*tank*</head> defences well to …" and "…high-sided <head>*tank*</head> , a **couple** of …", the context words "series" and "couple" belong to the same cluster, and thus our approach mistakenly assigned the same sense to the target word "tank" in these two sentences. Most of the other cases were caused by the errors of the automatic clustering algorithm and the noise in the web data.

V. Related Work

Several recent studies have stressed the benefits of using unsupervised word or phrase clustering as additional knowledge to improve supervised learning. For example, Miller et al. [13] proved that word clusters can significantly improve English name tagging. Ji [14] used cross-lingual predicate cluster acquisition to improve bilingual event extraction in an inductive learning framework. Lin and Wu [9] applied web-scale phrase clustering algorithm to improve name tagging and query classification. Pantel and Lin [15] described a clustering by committee algorithm to automatically discover word senses.

VI. Conclusions

We described a new approach of using automatic web-scale phrase clustering to improve word sense disambiguation. We demonstrated that without using any additional labeled data, this approach can capture some characteristics of the context matching required for WSD and thus significantly improved the performance over a typical baseline system using WordNet. In the future we are interested in investigating reliable clustering confidence metrics so that we can avoid noise in the web data. We will also attempt extending this approach to other more fine-grained sense disambiguation task such as event tagging.

Acknowledgment

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. NSF CAREER Award under Grant IIS-0953149, Google, Inc., DARPA GALE Program, CUNY Research Enhancement Program, PSC-CUNY Research Program, Faculty Publication Program and GRTI Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] David Vickrey, Luke Biewald, Marc Teyssier and Daphne Koller. 2005. Word-Sense Disambiguation for Machine Translation. *Proc. HLT/EMNLP 2005*.

[2] Dekai Wu. 2008. WSD for Semantic SMT: Phrase Sense Disambiguation. *Second Symposium on Innovations in Machine Translation Technologies (IMTT-2008)*.

[3] Ying Liu, Peter Scheuermann, Xingsen Li and Xingquan Zhu. 2007. Using WordNet to Disambiguate Word Senses for Text Classification. *Workshop on Text Data Mining in conjunction with 7th International conference on Computational Science*.

[4] David Yarowsky. 1992. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. *Proc. COLING 1992*.

[5] Alberto J. Canas, Alejandro Valerio, Juan Lalinde-Pulido, Marco Carvalho and Marco Arguedas. 2003. Using WordNet for Word Sense Disambiguation to Support Concept Map Construction. *LNCS 2857*, Springer-Berlin. 350-359.

[6] Hee-Cheoi Seo, Hoojung Chung, Hae-Chang Rim, Sung Hyon Myaeng and Soo-Hong Kim. 2004. Unsupervised Word Sense Disambiguation Using WordNet Relatives. *Computer Speech & Language*. Volume 18, Issue 3, pp. 253-273.

[7] Samuel Brody and Mirella Lapata. 2008. Good Neighbors Make Good Senses: Exploiting Distributional Similarity for Unsupervised WSD. *Proc. COLING 2008*.

[8] Christiane Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

[9] Dekang Lin and Xiaoyun Wu. 2009. Phrase Clustering for Discriminative Learning. *Proc. ACL 2009*.

[10] Dekang Lin, Ken Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani and Sushant Narsale. 2010. New Data, Tags and Tools for Web-Scale N-grams. Proc. LREC 2010.

[11] Zellig Harris. 1985. Distributional structure. In: Katz, J. J. (ed.) The Philosophy of Linguistics. New York: Oxford University Press. pp.26-47.

[12] Rada Mihalcea. The role of non-ambiguous words in natural language disambiguation. *Proc. RANLP 2003.*

[13] Scott Miller, Jethran Guinness and Alex Zamanian.2004. Name Tagging with Word Clusters and Discriminative Training. *Proc. HLT-NAACL2004*. pp. 337-342. Boston, USA.

[14] Heng Ji. 2009. Unsupervised Cross-lingual Predicate Cluster Acquisition to Improve Bi-lingual Event Extraction. *Proc. HLT-NAACL 2009 Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics.*

[15] Patrick Pantel and Dekang Lin. Automatically Discovered Word Senses. *Proc. HLT-NAACL 2003*.