Using the Web to Disambiguate Acronyms

Eiichiro Sumita[1, 2]
[1] NiCT
[2] ATR SLC
Kyoto 619-0288, JAPAN
eiichiro.sumita@atr.jp

Fumiaki Sugaya[3]
[3] KDDI R&D Labs
Saitama 356-8502, JAPAN
fsugaya@kddilabs.jp

Abstract

This paper proposes an automatic method for disambiguating an acronym with multiple definitions, considering the context surrounding the acronym. First, the method obtains the Web pages that include both the *acronym* and its *definitions*. Second, the method feeds them to the machine learner. Cross-validation tests results indicate that the current accuracy of obtaining the appropriate definition for an acronym is around 92% for two ambiguous definitions and around 86% for five ambiguous definitions.

1 Introduction

Acronyms are short forms of multiword expressions (we call them *definitions*) that are very convenient and commonly used, and are constantly invented independently everywhere. What each one stands for, however, is often ambiguous. For example, "ACL" has many different definitions, including "Anterior Cruciate Ligament (an injury)," "Access Control List (a concept in computer security)," and "Association for Computational Linguistics (an academic society)." People tend to write acronyms without their definition added nearby (**Table 1**), because acronyms are used to avoid the need to type long expressions. Consequently, there is a strong need to disambiguate acronyms in order to correctly analyze or retrieve text. It is crucial to recognize the correct acronym definition in information retrieval such as a blog search. Moreover, we need to know the meaning of an acronym to translate it correctly. To the best of our knowledge, no other studies have approached this problem.

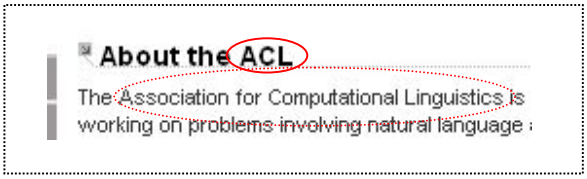

Figure 1 Acronyms and their definitions co-occur in some pages of the Web

On the other side of the coin, an acronym should be defined in its neighborhood. For instance, one may find pages that include a certain *acronym* and its *definition* (**Figure 1**).

First, our proposed method obtains Web pages that include both an *acronym* and its *definitions*. Second, the method feeds them to the machine learner, and the classification program can determine the correct definition according to the context information around the acronym in question.

Definition 1	**Anterior Cruciate Ligament**	**http://www.ehealthmd.com/library/aclteears**
She ended up with a torn **ACL**, MCL and did some other damage to her knee. (http://aphotofreak.blogspot.com/2006/01/ill-give-you-everything-i-have-good.html)		
Definition 2	**Access Control List**	**http://en.wikipedia.org/wiki**
Calculating a user's effective permissions requires more than simply looking up that user's name in the **ACL**. (http://www.mcsa-exam.com/2006/02/02/effective-permissions.html)		
Definition 3	**Association for Computational Linguistics**	**http://www.aclweb.org/**
It will be published in the upcoming leading **ACL** conference. (http://pahendra.blogspot.com/2005/06/june-14th.html)		

Table 1 Acronym "ACL" without its definition in three different meanings found in blogs

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 161–164,
New York, June 2006.

Here, we assume that the list of possible definitions for an acronym is given from sources external to this work. Listing pairs of acronyms and their original definitions, on which many studies have been done, such as Nadeau and Turney (2005), results in high performance. Some sites such as http://www.acronymsearch.com/ or http://www.findacronym.com/ provide us with this function.

This paper is arranged as follows. Section 2 explains our solution to the problem, and Section 3 reports experimental results. In Sections 4 and 5 we follow with some discussions and related works, and the paper concludes in Section 6.

2 The proposal

The idea behind this proposal is based on the observation that an *acronym* **often** co-occurs with its *definition* within a single Web page (**Figure 1**). For example, the acronym ACL co-occurs with one of its definitions, "Association for Computational Linguistics," **211,000 times** according to google.com.

Our proposal is a kind of word-sense disambiguation (Pedersen and Mihalcea, 2005). The hit pages can provide us with training data for disambiguating the acronym in question, and the snippets in the pages are fed into the learner of a classifier. Features used in classification will be explained in the latter half of this subsection.

We do not stick to a certain method of machine learning; any state-of-the-art method will suffice. In this paper we employed the decision-tree learning program provided in the WEKA project.

Collecting the training data from the Web

Our input is the acronym in question, A, and the set of its definitions, $\{D_k | k=1\sim K\}$.

```
for all k =1~K do
1.Search the Web using query of
   "A AND Dk."
2.Obtain the set of snippets, {Sl
   (A, Dk)| l=1~L}.
3.Separate Dk from Sl and obtain
   the set of training
   data,{(Tl(A), Dk)| l=1~L}.
  End
```

In the experiment, **L is set to 1,000**. Thus, we have for each definition D_k of A, at most 1,000 training data.

Training the classifier

From training data $T_l(A)$, we create feature vectors, which are fed into the learner of the decision tree with correct definition D_k for the acronym A.

Here, we write $T_l(A)$ as $W_{-m}\ W_{-(m-1)}$ ... $W_{-2}\ W_{-1}$ A $W_1\ W_2$ ... $W_{m-1}\ W_m$, where m is from 2 to M, which is called the window size hereafter.

We use keywords within the window of the snippet as features, which are binary, i.e., if the keyword exists in $T_l(A)$, then it is true. Otherwise, it is null.

Keywords are defined in this experiment as the top N frequent words [1], but for A in the bag consisting of all words in $\{T_l(A)\}$. For example, keywords for "ACL" are "*Air, Control, and, Advanced, Agents, MS, Computational, Akumiitti, Cruciate, org, of, CMOS, Language, BOS, Agent, gt, HTML, Meeting, with, html, Linguistics, List, Active, EOS, USA, is, access, Adobe, ACL, ACM, BETA, Manager, list, Proceedings, In, A, League, knee, Anterior, ligament, injuries, reconstruction, injury, on, The, tears, tear, control, as, a, Injury, lt, for, Annual, Association, Access, An, that, this, may, an, you, quot, in, the, one, can, This, by, or, be, to, Logic, 39, are, has, 1, from, middot.*"

3 Experiment

3.1 Acronym and definition preparation

We downloaded a list of acronyms in capital letters only from *Wikipedia* and filtered them by eliminating acronyms shorter than three letters. Then we obtained definitions for each acronym from http://www.acronymsearch.com/ and discarded acronyms that have less than five definitions. Finally, we randomly selected 20 acronyms.

We now have 20 typical acronyms whose ambiguity is more than or equal to five. For each acronym A, a list of definitions $\{ D_k \mid k=1\sim K\ K>=5 \}$, whose elements are ordered by the count of page including A and D_k, is used for the experiment.

[1] In this paper, **N is set to 100**.

3.2 Ambiguity and accuracy

Here we examine the relationship between the degree of ambiguity and classification accuracy by using a cross-validation test for the training data.

#Class	M=2	M=5	M=10	Base
2	**88.7%**	**90.1%**	**92.4%**	82.3%

Table 2 Ambiguity of two

#Class	M=2	M=5	M=10	Base
5	**78.6%**	**82.6%**	**86.0%**	76.5%

Table 3 Ambiguity of five

Ambiguity of two

The first experiment was performed with the selected twenty acronyms by limiting the top two most frequent definitions. **Table 2** summarizes the ten-fold cross validation. While the accuracy changes acronym by acronym, the average is high about 90% of the time. The M in the table denotes the window size, and the longer the window, the higher the accuracy.

The "base" column displays the average accuracy of the baseline method that always picks the most frequent definition. The proposed method achieves better accuracy than the baseline.

Ambiguity of five

Next, we move on to the ambiguity of five (**Table 3**). As expected, the performance is poorer than the abovementioned case, though it is still high, i.e., the average is about 80%. Other than this, our observations were similar to those for the ambiguity of two.

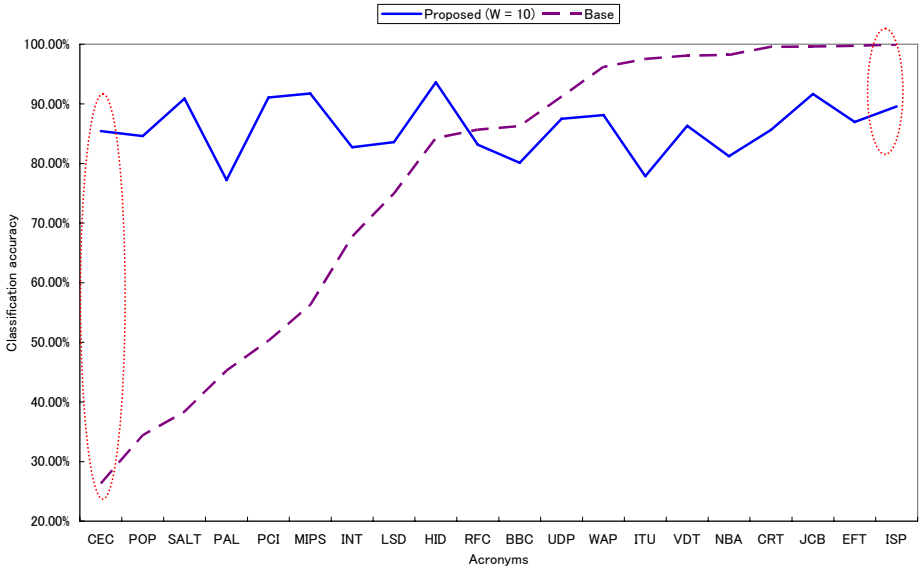

Figure 2 Bias in distribution of definitions (ambiguity of 5)

4 Discussion on biased data

4.1 Problem caused by biased distribution and a countermeasure against it

For some words, the baseline is more accurate than the proposed method because the baseline method reaches all occurrences on the Web thanks to the search engine, whereas our method limits the number of training data by L as mentioned in Section 2. The average quantity of training data was about 830 due to the limit of L, 1,000. The distribution of these training data is rather flat. This causes our classifier to fail in some cases. For example, for the acronym "ISP," the most frequent definition out of five has a share of 99.9% (**Table 4**) on the Web, whereas the distribution in the training data is different from the sharp distribution. Thus, our classification accuracy is not as good as that of the baseline.

Considering the acronym "CEC," the most frequent out of five definitions has the much smaller share of 26.3% on the Web (**Table 5**), whereas the

distribution in the training data is similar to the flat distribution of real data. Furthermore, the decision tree learns the classification well, whereas the baseline method performs terribly.

These two extreme cases indicate that for some acronyms, our proposed method is beaten by the baseline method. The slanting line in **Figure 2** shows the baseline performance compared with our proposed method. In the case where our method is strong, the gain is large, and where our method is weak, the reduction is relatively small. The average performance of our proposed method is higher than that of the baseline.

Definition	Page hits
Internet Service Provider	3,590,000
International Standardized Profile	776
Integrated Support Plan	474
Interactive String Processor	287
Integrated System Peripheral control	266

Table 4 Sharp distribution for "ISP"

Definition	Page hits
California Energy Commission	161,000
Council for Exceptional Children	159,000
Commission of the European Communities	138,000
Commission for Environmental Cooperation	77,400
Cation Exchange Capacity	76,400

Table 5 Flat distribution for "CEC"

A possible countermeasure to this problem would be to incorporate prior probability into the learning process.

4.2 Possible dissimilarity of training and real data

The training data used in the above experiment were only the type of snippets that contain **acronyms and their definitions**; there is no guarantee for documents that contain only **acronyms** are similar to the training data. Therefore, learning is not necessarily successful for real data. However, we tested our algorithm for a similar problem introduced in Section 5.1, where we conducted an open test and found a promising result, suggesting that the above-mentioned fear is groundless.

5 Related works

5.1 Reading proper names

The contribution of this paper is to propose a method to use Web pages for a disambiguation task. The method is applicable to different problems such as reading Japanese proper names (Sumita and Sugaya, 2006). Using a Web page containing a name and its syllabary, it is possible to learn how to read proper names with multiple readings in a similar way. The accuracy in our experiment was around 90% for open data.

5.2 The Web as a corpus

Recently, the Web has been used as a corpus in the NLP community, where mainly counts of hit pages have been exploited (Kilgarriff and Grefenstette, 2003). However, our proposal, Web-Based Language Modeling (Sarikaya, 2005), and Bootstrapping Large Sense-Tagged corpora (Mihalcea, 2002) use the content within the hit pages.

6 Conclusion

This paper proposed an automatic method of disambiguating an acronym with multiple definitions, considering the context. First, the method obtains the Web pages that include both the acronym and its definitions. Second, the method feeds them to the learner for classification. Cross-validation test results obtained to date indicate that the accuracy of obtaining the most appropriate definition for an acronym is around 92% for two ambiguous definitions and around 86% for five ambiguous definitions.

References

A. Kilgarriff and G. Grefenstette. 2003. "Introduction to the special issue on the Web as a corpus," Computational Linguistics 29(3): 333-348.

Rada. F. Mihalcea, 2002. "Bootstrapping Large Sense-Tagged Corpora," Proc. of LREC, pp. 1407-1411.

David Nadeau and Peter D. Turney, 2005. "A supervised learning approach to acronym identification," 18th Canadian Conference on Artificial Intelligence, LNAI3501.

Ted Pedersen and Rada. F. Mihalcea, "Advances in Word Sense Disambiguation," tutorial at ACL 2005. http://www.d.umn.edu/~tpederse/WSDTutorial.html.

Ruhi Sarikaya, Hong-kwang Jeff Kuo, and Yuqing Gao, 2005. Impact of Web-Based Language Modeling on Speech Understanding, Proc. of ASRU, pp. 268-271.

Eiichiro Sumita and Fumiaki Sugaya,. 2006. "Word Pronunciation Disambiguation using the Web," Proc. of HLT-NAACL 2006.