KUNLP system using Classification Information Model at SENSEVAL-2

Hee-Cheol Seo, Sang-Zoo Lee, Hae-Chang Rim
Dept. of Computer Science and Engineering,
Korea University
1, 5-ka, Anam-dong
Seongbuk-Gu, Seoul, 136-701, Korea
{hcseo,zoo,rim}@nlp.korea.ac.kr

Ho Lee
Astronest Inc.
135-090 3rd floor, Hanam BD
157-18 Samsung-Dong
Kangnam-Gu, Seoul, Korea
leeho@astronest.com

Abstract

The classification information model or CIM classifies instances by considering the discrimination ability of their features, which was proven to be useful for word sense disambiguation at SENSEVAL-1. But the CIM has a problem of information loss. KUNLP system at SENSEVAL-2 uses a modified version of the CIM for word sense disambiguation.

We used three types of features for word sense disambiguation: local, topical, and bigram context. Local and topical context are similar to Chodorow's context and refer to only unigram information. The window of a bigram context is similar to that of a local context but a bigram context refers to only bigram information.

We participated in the English lexical sample task and the Korean lexical sample task, where our systems ranked high.

1 Introduction

The classification information model(Ho, 1997) is the model that classifies instances by considering the discrimination ability of their features. In the CIM, a feature with high discrimination ability contributes to the classification more than one with low discrimination ability. Hence, we can omit the feature selection procedure.

The CIM has a kind of information loss problem due to the assumption that a feature contributes to only one class. We devised a modified version of the CIM where a feature can contribute to all classes.

Word sense disambiguation task can be treated as a kind of classification process(Ho, 2000). When a classification technique is applied to word sense disambiguation, an instance corresponds to a context containing a polysemous word and its class to the proper sense of the word, and one of its features to a piece of context information. As a classification problem, word sense disambiguation task can be solved by the CIM.

We used three types of features for word sense disambiguation: local, topical, and bigram context. Local and topical context are similar to Chodorow's context(Chodorow, 2000) and consist of only unigram information. A bigram context has a similar window to a local context but consists of only bigram information.

2 KUNLP system

To disambiguate senses, we did two phases: corpus preprocessing and sense disambiguation. Figure 1 shows the flow chart of our system.

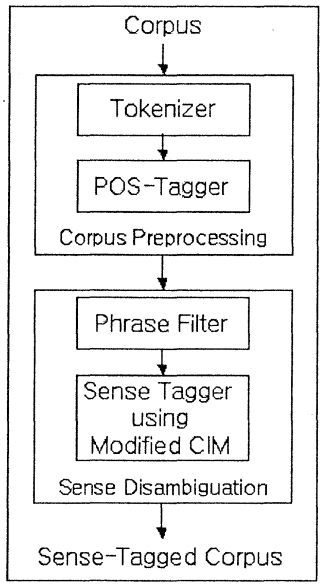

Figure 1: Flow chart of KUNLP system

2.1 Corpus preprocessing

At the corpus preprocessing phase, we tokenized a corpus and then tagged it with parts-of-speech using Brill's Tagger(Brill, 1994). The tokenizer just separates symbols from a word. For example, a sentence "*I'm straight, white, no longer middle class, anti-IRA, have ...*" is tokenized to "*I 'm stright , white , no longer middle class , anti - IRA , have ...*". Unlike other symbols, an apostrophe is not separated from the following characters.

2.2 Phrase filtering

At the phrase filtering phase, we filtered senses using the satellite feature, which is marked with *sat* tag in training and test corpus given by the task organizer. For example, in a sentence *This air of disengagement <head sats="carry_over.067:0">carried</head> <sat id="carry_over.067:0">over</sat> to his apparent attitude toward his things, carried over* is a phrase and also a satellite feature.

Phrase filtering is applied to sense disambiguation as in Table 1

Table 1: phrase filtering and sense disambiguation

if the number of filtered senses = 1 then determine sense
else if the number of filtered senses > 1 then execute sense-tagger with the filtered senses
else if the number of filtered senses = 0 then execute sense-tagger with all senses

There are satellite features in the English lexical sample, but not in the Korean lexical sample. Hence, phrase filtering was applied only in the English lexical sample task.

2.3 Classification Information Model (CIM)

The CIM is a kind of classification model based on the entropy theory. Given an input instance, the CIM decides the proper class of the instance by considering individual decisions made by each feature of the instance. In the model, the proper class of an instance,X, is determined by Equation 1.

$$Class(X) \stackrel{\text{def}}{=} \arg\max_{class_j} \text{Rel}(class_j, X) \quad (1)$$

where $class_j$ is the j-th class and $\text{Rel}(class_j, X)$ is the relevance between the j-th class and the instance X. Here, if we assume that features are independent of each other, the relevance can be defined as in Equation 2.

$$\text{Rel}(class_j, X) = \sum_{i=1}^{m} x_i w_{ij} \quad (2)$$

where m is the size of the feature set, x_i is the value of the i-th feature and w_{ij} is the weight of the i-th feature for the j-th class. In Equation 2, x_i has a binary value (1 if the feature occurs within the window, 0 otherwise) and w_{ij} is defined in terms of classification information.

The classification information of a feature is composed of two components. One is the discrimination score (DS), which represents the discrimination ability of classifying instances. The other is the most probable class (MPC), which represents the most closely related class to the feature. w_{ij} is defined by using these two components as follows:

$$w_{ij} \stackrel{\text{def}}{=} \begin{cases} \text{DS}_i & \text{if } class_j = \text{MPC}_i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

In Equation 3, DS_i and MPC_i represent the DS and MPC of the i-th feature, respectively. In the CIM, DS and MPC are defined in terms of the conditional probability of a class given a feature, which is normalized by the corpus size. The normalized conditional probability is defined as follows:

$$\begin{aligned} \hat{p}_{ji} &\stackrel{\text{def}}{=} \frac{p(class_j|f_i)\frac{\overline{N(class)}}{N(class_j)}}{\sum_{k=1}^{n} p(class_k|f_i)\frac{\overline{N(class)}}{N(class_k)}} \\ &= \frac{p(f_i|class_j)}{\sum_{k=1}^{n} p(f_i|class_k)} \end{aligned} \quad (4)$$

In Equation 4, $\hat{p}_{ji}$ is a normalized conditional probability, $N(class_j)$ is the number of instances belonging to the j-th class in the training data, $\overline{N(class)}$ is the average number of instances for each class and n is the number of classes. Given the normalized conditional probability distribution, DSs and MPCs are defined as follows:

$$\begin{aligned} DS_i &\stackrel{\text{def}}{=} \log_2 n - H(\hat{p}_i) \\ &= \log_2 n + \sum_{j=1}^{n} \hat{p}_{ji} \log_2 \hat{p}_{ji} \end{aligned} \quad (5)$$

$$\begin{aligned} MPC_i &\stackrel{\text{def}}{=} \arg\max_{class_j} \hat{p}_{ji} \\ &= \arg\max_{class_j} p(f_i|class_j) \end{aligned} \quad (6)$$

In Equation 5, $H(\hat{p}_i)$ is the entropy of the i-th feature over the normalized conditional probability distribution.

2.4 Modifying CIM

The CIM has a problem caused by using MPCs, which is information loss. For example, let us consider the situation in Table 2 and Table 3. Table 2 shows the normalized conditional probability distribution, DSs and MPCs of features in an instance. Table 3 shows the weights and the relevance values at the CIM using w_{ij} and at the modified CIM using $\hat{w}_{ij}$, for the instance of Table 2. The feature f_1 co-occurred with $class_1$ and $class_2$ and the MPC of f_1 is $class_1$ at Table 2. In the CIM, this feature

Table 2: A normalized conditional probability, DSs and MPCs of features of an instance

feature	normalized conditional probability($\hat{p}_{ji}$)				DS	MPC
	$class_1$	$class_2$	$class_3$	$class_4$		
f_1	0.7	0.3	0	0	1.1187	$class_1$
f_2	0	0.4	0.6	0	1.0290	$class_3$
f_3	0	0.4	0.1	0.5	0.6390	$class_4$

Table 3: The weights and the relevance values at the CIM using w_{ij} and at the modified CIM using $\hat{w}_{ij}$, for the instance of Table 2

feature	weight(w_{ij})				weight($\hat{w}_{ij}$)			
	$class_1$	$class_2$	$class_3$	$class_4$	$class_1$	$class_2$	$class_3$	$class_4$
f_1	1.1187	0	0	0	0.7831	0.3356	0	0
f_2	0	0	1.0290	0	0	0.4116	0.6174	0
f_3	0	0	0	0.6390	0	0.2556	0.0639	0.3195
$Rel(class_j, X)$	1.1187	0	1.0290	0.6390	0.7831	1.0028	0.6813	0.3195

contributes to only $class_1$. Actually the feature f_1 can contribute to distinguishing $class_2$ from $class_3$ if it consults the normalized conditional probability distribution. In the CIM, however, the feature can not distinguish them because their weights have the same value.

Another aspect of the problem is that the CIM fails to capture the minor contribution of features, which is crucial in the case where the sum of the minor contribution of features to a non-MPC class dominates that of the major contribution of features to MPC classes. For example, at Table 2, all features, f_1, f_2, and f_3, have different MPCs: $class_1$, $class_3$ and $class_4$, respectively. it is also obvious that they have some minor contribution to the $class_2$. The CIM will classify the instance as $class_1$ because $Rel(class_1, X) = 1.1187$ is the maximum number among the $Rel(class_j, X)$. However, if we consider the minor contribution of all the features, we prefer $class_2$ to $class_1$ because $class_2$ intuitively gains the total contribution more than $class_1$.

A solution to the problem may be not to use MPCs, but to use a measure of contribution of a feature to a class which is proportional to the discrimination score of the feaure and the normalized conditional probability of the class given the feature. The modified CIM can be defined as follows:

$$\mathrm{Rel}(class_j, X) = \sum_{i=1}^{m} x_i \hat{w}_{ij} \quad (7)$$

$$\hat{w}_{ij} \stackrel{\mathrm{def}}{=} DS_i \times \hat{p}_{ji} \quad (8)$$

As shown in Table 3, the $\hat{w}_{12}$ is larger than $\hat{w}_{13}(0.3356 > 0)$ and the instance is classified not as $class_1$ but as $class_2$ because $Rel(class_2, X) = 1.0028 > Rel(class_1, X) = 0.7831$, which is based on the modified CIM.

2.5 Feature Space

We used three types of features for word sense disambiguation: local, topical and bigram context. In the preliminary experiment, we have observed that, when the CIM considered all these three types of features, it mostly achieved the best result.

2.5.1 Local context

In a local context, there can be features of the following templates for all words within its window:

- in the English lexical sample task
  - *word_position* : a word and its position
  - *word_POS* : a word and its part-of-speech
  - *POS_position* : the part-of-speech and position of a word
- in the Korean lexical sample task
  - *morpheme_position* : a morpheme[1] and its position.
  - *morpheme_POS* : a morpheme and its part-of-speech.
  - *POS_position* : the part-of-speech and position of a morpheme

In the English lexical sample task, *word* is a surface form and can be either one of open-class words whose POS is one of the noun, verb, adjective, and adverb; or one of closed-class words whose POS is

[1] A Korean sentence is composed of one or more *eojeols*, which are separated by spaces, and an *eojeol* consists of one or more morphemes.

one of the determiner, preposition, pronoun, and punctuation. The window size of ±3 words in the English lexical sample task and the window size from −2 to +3 word in the Korean lexical sample task were empirically chosen.

In the first phase of the experiments, we used just one complicated template, *word_position_POS*(in Korean *morpheme_position_POS*), which brought about data sparseness problem. So we split the template into three simpler templates.

2.5.2 Topical context

A topical context includes features of the following templates for all open-class words within its window:

- in the English lexical sample task
  - *word* : an open-class word.
- in the Korean lexical sample task
  - *morpheme* : an open-class morpheme.

The window size of ±1 sentences in the English lexical sample task and the window size of all sentences in the Korean lexical sample task were empirically chosen.

2.5.3 Bigram context

In a bigram context, there can be features of the following templates for all word-pairs within its window:

- in the English lexical sample task
  - $(word_i, word_j)$: the i-th word and j-th word $(i>j)$
  - $(word_i, POS_j)$: the i-th word and j-th part-of-speech $(i>j)$
- in the Korean lexical sample task
  - $(eojeol_i, eojeol_j)$: the i-th eojeol and j-th eojeol $(i>j)$

Unlike local and topical contexts, bigram contexts are composed of only bigram information surrounding the polysemous word. The window size of ±2 words in the English lexical sample task and the window size from −2 to +3 word in the Korean lexical sample task were empirically chosen.

3 Experimental Result

The following tables show the results of our systems at SENSEVAL-2 (Table 4). For the Korean lexical sample task at SENSEVAL-2, only fine-grained sense distinction was made.

Table 4: Results of KUNLP systems at SENSEVAL-2

task	prec.	recall
English Lexical Sample (fine g.)	0.629	0.629
English Lexical Sample (coarse g.)	0.697	0.697
Korean Lexical Sample (fine g.)	0.698	0.74

4 Conclusion

We have described the modified CIM used for word sense disambiguation at SENSEVAL-2. In the experiments, three types of features; local, topical, and bigram context, are used. Our system ranked as the highest at the Korean lexical sample task and as the topmost group at the English lexical sample task among the supervised models at SENSEVAL-2. Consequently, the results back up the fact that the modified CIM and three types of features are useful for discriminating word senses.

References

Eric Brill 1994. Some advances in rule-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94).*

Martin Chodorow, Claudia Leacock and George A. Miller 2000. A Topical/Local Classifier for Word Sense Identification. In *Computers and the Humanities 34: 115-120.*

Ho Lee, Dae-Ho Baek and Hae-Chang Rim 1997. Word Sense Disambiguation Based on The Information Theory. In *Proceedings of Research on Computational Linguisitcs Conference.*

Ho Lee, Hae-Chang Rim and JungYun Seo 2000. Word Sense Disambiguation Using the Classification Information Model. In *Computers and the Humanities 34: 141-146.*