The Effect of Answer Patterns for Supervised Named Entity Recognition in Thai *

Nutcha Tirasaroj and Wirote Aroonmanakun

Department of Linguistics, Chulalongkorn University,
Bangkok, Thailand
pernn39@gmail.com, awirote@chula.ac.th

Abstract. In this paper, we present Thai named entity recognition (NER) systems using supervised Conditional Random Fields (CRFs) with various answer patterns to find out whether different answer patterns would affect the performance of the systems. Every system used the same set of features except the answers in the training corpus. There are 5 patterns of answer used in this study. The results show that the answer tags having more information can help improve the system performance than those having less information.

Keywords: Named entity recognition, Conditional Random Fields, Supervised machine learning, Thai language processing

Introduction

Named entity recognition (NER) was first introduced in MUC-6 in 1990's. According to MUC, NER is divided into 3 main groups. The first one is entity names including person, organization, and location names. The second one is temporal expressions and the last one is numerical expressions which are monetary values and percentages. Among these 3 groups, the first one is the most difficult to be recognized because the structures of the names are complicated and the contexts they occur are varying. As a result, a number of research on NER as well as this study have focused on the first group.

Most research on NER using statistical models mainly focused on feature selection since features are the most important part of the models and directly affect the performance of the systems. It was assumed that the answer given to the models has no effect on the system performance. However, in Chinese, some research on NER (Mao *et al.*, 2008; Yu *et al.*, 2008), the work of Yu *et al*. (2008) proved that the number of answer tags can help improve the system performance. The use of more tags yields better results because it gives more information of entity names' boundaries to the system for learning. For machine learning, it tends to be that the more information is provided, the better the performance of the system is. Research on Thai NER using statistical models such as SVMs (Suwanno *et al.*, 2007), HMM (Sansing and Prayot, 2010), ME (Chanlekha and Kawtrakul, 2004), CRFs (Tepdang *et al.*; Tirasaroj and Aroonmanakun, 2009), etc. also focused on feature selections. None has studied the effect of answer patterns for their supervised systems. The answer pattern is the way we give the answer in the training corpus for supervised machine learning system. The answer suggests how much information of NEs provided for the systems. For example, the answer can be only the information of NE types or it can include the NE boundaries. In addition, most of previous

* The work reported in this paper was supported by The Thailand Research Fund (TRF) under grant no MSG53Z0008, and partially supported by Chulalongkorn University Centenary Academic Development Project. Moreover, we would like to thank NECTEC for their BEST 2009 corpus and reviewers for their suggestions and comments.

Copyright 2011 by Nutcha Tirasaroj and Wirote Aroonmanakun

research did not distinguish metonymic names, which are location names referring to organizations and organization names referring to locations.

Thus, apart from presenting the NER systems, this study is conducted to find out if the more informative answers can improve the system performance of Thai NER like in Chinese or not. In addition, we will find out the differences of results when metonymic names are marked. Our NER systems are based on CRF models. Recently, there was a comparative study on Thai word segmentation approaches (Haruechaiyasak *et al.*, 2008). The results showed that CRFs yielded the best performance when compared with other machine learning methods, Naïve Bayes, decision tree, and Support Vector Machine. CRFs were also applied to Thai NER (Tepdang *et al.*; Tirasaroj and Aroonmanakun, 2009) and yielded about 80% of recognition rate or F-measure. In this study, we used CRF++0.53 implemented by Taku Kudo.

Related Work

There are a number of research focusing on Thai named entity recognition. Charoenpornsawat *et al*. (1998) used feature-based approach such as context words, collocations, and heuristic rules to extract the candidates and then used Winnow algorithm to recognize the entity names. The accuracy was 92.17%. Chanlekha *et al*. (2002) applied statistical and heuristic rule-based model to POS tagged data. The results showed that the model did have much problem when extract the names from magazines but when the model was applied to the newspapers, the accuracy was quite low. It was due to the written style of magazines that the names were usually in regular patterns.

Tirasaroj and Aroonmanakun (2009) compared the performance of the systems between word-based and syllable-based system by applying CRF. The results showed no significant difference. Although in their work, they mentioned about the NE tags that differentiated between common named entities – person, organization, and location names - and metonymic names, the organization names referring to the locations and vice versa. Their answers did not divided into five categories like their tags but just three categories as other previous research.

In Thai NER, the researchers have never mentioned about their answer patterns used in their systems but in Chinese, there are some research proving that the number of tag types does have an effect on supervised learning systems. The amount of tags used in 2008 is more than those used in 2006. In 2006, Feng *et al*. (2006) used BIO tags; B for the beginning of names, I for the inside of names, and O for others. In 2008, Mao *et al*. (2008) used BIOE and Yu *et al*. (2008) used BIOES instead of BIO. They added E and S; E for the end of names and S for single character names. From their research, giving more tags could help improve the system performance. In our study, we will apply NE type tags like in Tirasaroj and Aroonmanakun's work and for NE boundaries, we will adapt from BIOE. We will explain in more detail in the section about patterns of answer.

Conditional Random Filed Models

Conditional Random Fields (CRFs) (Lafferty *et al*., 2001), which are developed from MEMMs, is a discriminative undirected graphical model. They are usually used for segmenting and labeling sequence data. CRF models the conditional distribution P(Y|X), where Y is the hidden label sequence or output and X is the observation or input. A linear chain CRF defines the conditional probability of sequence Y when given the observation sequence X as follows:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \quad (1)$$

Z is a normalization factor defined as:

$$Z(x) = \sum_{y} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \quad (2)$$

where $f_k(y_{t-1}, y_t, x, t))$ is an arbitrary feature function, λ_k is a learned weight for each feature function.

Corpus

The corpus used in this study is taken from BEST 2009 corpus. BEST 2009 is the corpus created for training and testing word segmentation systems in "BEST 2009: THAI WORD SEGMENTATION SOFTWARE CONTEST." Many genres of data were collected in BEST but only news genre was used in our study because in the news there are all three kinds of named entities- person, organization, and location names- and the number of entities are much more various than those found in other genres like fictions and encyclopedia. Thai word segmentation program (Aroonmanakun, 2002) was used to segment all data in our corpus. The number of tokens is about 367,600 tokens.

As we applied supervised learning system, we tagged all answers in the corpus. The data from BEST 2009 had been tagged the NE boundaries but not NE types. We then had to add NE types manually. As one of the main purposes of this study was to compare the answer patterns of the systems, we tried to give as much information of answers as possible. We followed Tirasaroj and Aroonmakun (2009) so that there were five NE types in our corpus. Apart from common NE types, person, organization, and location names, there were two more tags for metonymic names, organization names referring to locations and vice versa. The followings were NE tags in our corpus:

1. <persName>…</persName> is for person names including first names, last names, nicknames, alias, etc. The titles and honorifics right before the names are also included such as

<persName>ลุงสมหมาย</persName>

Uncle Sommai

2. <orgName>…</orgName> is for organization names. In case that the full names are followed by their abbreviated names, they are tagged separately like

<orgName>กระทรวงสาธารณสุข</orgName><orgName><abb>สธ.</abb></orgName>

Ministry of Public Health MOPH

3. <placeName>…</placeName> is for location names such as

ที่<placeName>นครปฐม</placeName>

at Nakornpathom

4. <orgName ref="loc">…</orgName> is for organization names used to refer to locations. Generally, if the locations do not have their specific names, their names would undoubtedly be the same as the organization situated in that place. For example,

เข้ามารักษาโรคที่<orgName ref="loc">โรงพยาบาลสัตว์</orgName>

had exam at Veterinary hospital

5. <placeName ref="org">…</placeName> is for location names used as organization names. Mostly, this kind of names is used to refer to an organization, especially official offices and the government such as

<placeName ref="org">ประเทศไทย</placeName>ได้แจ้งไปยังองค์การระหว่างประเทศเรียบร้อยแล้ว

Thailand has already informed International Organizations

Besides NE tags, there is a tag for abbreviations. The tag <abb>…</abb> is used to tag every abbreviation in the corpus. In case of the abbreviations like titles or honorifics which are parts

of names, the abbreviation tags would be tagged inside NE tags like the example in the organization tag above.

Features and Patterns of Answer

In this study, we applied both open and closed features. The open features are the knowledge we got outside the corpus while the closed features are those extracted from the training and testing data. There are 7 features used in this study; two open features, three closed features, and two baseline features. The open and closed features are all binary features. In this study, the deep linguistic features such as POS, functional features were not used because this is the preliminary stage of our study and these features take a lot of time in preparing the training corpus. The followings are the details of each feature used in this study.

The open features: applying only closed features may not be enough for the system as the entity names sometimes occur in ambiguous context or without any clue words. The gazetteers were used to help extract this kind of names. Our gazetteers include the lists of person, organization, and location names, abbreviated organization names, and personal titles and honorifics. If the token matched with any part of the names in the list, this feature would be set to 'Y'.

Another outside knowledge is a general word list. This feature was set from the patterns of names that the names can composed of known and unknown strings. However, although those known strings are words found in dictionary, they are sometimes not words used in daily life. For example, "อัปสร" (Upsorn) is a synonym of 'นางฟ้า' (Nang-faa); both mean an angel, but no one generally calls the angel as "อัปสร" except in literature. Then, words that are not frequently found in daily life are possibly a name or a part of the name. So, in this case, the feature for the token not found in the list was set to 'Y'.

The closed features: there were three closed features we got from the corpus, including abbreviation, context clues, and repeated NE. The details of each feature are as follows:

The abbreviation feature was considered from the characteristic of Thai entity names that they are usually preceded by the abbreviation. Person names usually co-occur with titles and honorifics such as น.ส. (Miss), ดร. (Ph.D), organization names with company designators such as บจ. (company limited), and location names with location indicators like ถ. (Road), จ. (province). Those abbreviations can help indicate not only named entities' boundaries but also their types. For this feature, the sequence of three tokens would be set to 'Y' if one of them was an abbreviation.

The context clues are the key factors in assigning the types of metonymic names. For example, the location names, in fact, could not co-occur with verbs since only persons and organizations can do some actions. Thus, whenever the location names co-occur with verbs, it implies that these names are used as metonymic names referring to the organization. The list of context clues was extracted from the three tokens before and after the entity names in the training corpus. If the token matched with any token in the list, the feature was set to 'Y'.

Repeated NE is set from our assumption that most named entities do not occur once in the text. Then, if the strings of words appeared several times in the text, we assumed that they would probably be entity names. For this feature, if the three tokens, including the current token and the tokens before and after, occurred together more than three times in the text, they would be set to 'Y'.

The baseline features: apart from the open and closed features, there were baseline features used as templates for running CRF. Our baseline features were unigram and bigram.

Besides feature selection, we have to set the patterns of answer. There are 5 answer patterns used in this study. Each pattern has different information of named entities' boundaries and types. The details are as shown in Table 1:

Table 1: Answer patterns

No	Pattern
1	P, O, L, X
2	B, I, X – PER, ORG, LOC
3	B, I, X – P, O, L, LO, OL
4	B, I, E, X – PER, ORG, LOC
5	B, I, E, X – P, O, L, LO, OL

Pattern 1 has the least information and Pattern 5 has the most information. Pattern 1 has no information of named entities' boundaries but only basic entity types. The entity boundaries are BIEX, B for the beginning of a name, I for a token inside a name, E for the end of a name, and X for others. The answer pattern 1, 2, and 4 have basic entity types. The metonymic names are included in their referred types. The pattern 3 and 5 have all 5 entity types.

An example of training data, including features and answer, is shown in figure 1.

กระทรวง	N	N ..	B-ORG
สาธารณสุข	N	N ..	E-ORG
<s>	N	N ..	X
องค์การ	N	N ..	B-ORG
อนามัย	N	Y ..	I-ORG
โลก	N	Y ..	E-ORG
และ	N	Y ..	X
…			

Figure 1: Example of training data with the 4th answer pattern

The first column is the token. From the second column to the one before the last column, they are features which are all binary features. The last column is the answer given to CRF. Every answer pattern uses the same set of features.

Experimental Results

In this experiment, we used 10-fold cross validation and evaluated our systems with precision (P), recall (R), and F-measure (F). Only exactly matched answer is counted as correct answer. Table 2 shows the experimental results of all systems.

The performances of the systems with answer pattern 2 to 5 are almost indifferent, the difference of the values of f-measure is less than 1%. However, when compared with pattern 1, their performances are about 3% better. From all of the answer patterns, f-measure of pattern 4 is the highest while pattern 1 is the lowest. The results show that the patterns containing more information tend to improve the systems' performances than those having less information.

Although pattern 5 has most information, it is not the one getting the best performance. The main reason may be that the training corpus is not large enough. Since the named entities are divided into 5 categories, the system, therefore, needs a large number of examples of names in each category for training. However, from all of the entity types, the organization names referring to the locations have the lowest number of examples (147 instances), while the highest number of entity names is 5,672 instances which are person names. As a result, the f-measures of this kind of names, both in pattern 3 and 5, are low when compared with other named entity types.

Table 2: Experimental results

NE	P(%)	R(%)	F(%)
1. P, O, L, X			
PER	85.69	83.93	84.77
ORG	77.28	70.55	73.75
LOC	76.82	70.57	73.52
ALL	80.39	75.26	77.73
2. B, I, X – PER, ORG, LOC			
PER	90.59	86.37	88.41
ORG	82.68	74.61	78.42
LOC	80.72	70.71	75.31
ALL	85.24	77.67	81.26
3. B, I, X – P, O, L, LO, OL			
P	90.14	86.17	88.08
O	82.57	75.15	78.65
L	80.44	73.26	76.55
LO	81.43	67.32	73.43
OL	75.39	43.71	54.89
ALL	84.82	77.09	80.75
4. B, I, E, X – PER, ORG, LOC			
PER	92.05	86.50	89.16
ORG	82.19	74.23	77.99
LOC	79.92	70.77	74.98
ALL	85.37	77.64	81.30
5. B, I, E, X – P, O, L, LO, OL			
P	91.52	86.47	88.89
O	82.67	75.38	78.83
L	80.04	73.35	76.45
LO	81.04	68.00	73.72
OL	77.35	43.57	55.35
ALL	85.29	77.37	81.12

Another main problem of metonymic names is that the systems are confused about common names and metonymic names. The influential factor that can differentiate them is the context in which they occur. For instance, if the location names are followed by verbs or the organization names are preceded by prepositions, they tend to be metonymic names rather than common names. However, in this study, we did not use the deep linguistic features like POS tagging, so the systems could not know the part of speech of each word. Moreover, one word can have more than one part of speech. As a result, considering only forms of words like in our work is not enough for recognizing the differences between common names and metonymic names. We believe that the problem of metonymic name classification may be one factor that more or less affects the accuracy of organization and location name recognition of many previous studies.

The main problem of pattern 1 is the boundaries of named entities. Although classifying the entities' types correctly, the system tended to merge 2 named entities that were separated by a space but in the same category into a name. The examples are as follows:

<persName>จอย</persName><s><persName>รินลณี</persName>

<placeName>ตะกั่วทุ่ง</placeName><s><placeName>ตะกั่วป่า</placeName>

Form the performance of all systems shown in table 2, we can see that the answer given to the system more or less affect the system performance. Many researchers give importance to feature selection and overlook the answer designed for the system. However, from this study, it indicates that every process is important to the system.

Discussion

When considered overall performance in table 2, we will see that the f-measures of person names in every system are much higher than others. Apart from the highest number of samples in the corpus, the person names usually appear in regular patterns such as occurring with the titles, the first names separated from the last names by spaces, etc. As a result, the systems could easily recognize the person names rather than other NE types. For the organization names, although they do not appear in the regular patterns like person names, the way they appear in the text seems like a pattern. The organization names that have abbreviated forms are usually first introduced with the full names in the text and later appear with their abbreviated names. Thus, the f-measures of organization names are better than those of location names in every system.

In addition to f-measures, the recalls were quite low. The main problem is that the systems could hardly extract the names appearing without clue words such as person names appearing without the titles or honorifics or the names that are in ambiguous context. For example, "ส่งเสริม" in "ผศ. น.สพ. ดร. ทวีศักดิ์ ส่งเสริม" (Assistant Prof. Dr. Taweesak Songserm) is in fact a last name but has the same form like a verb "ส่งเสริม", which means 'support' in Thai. As a result, the system could extract only "ผศ. น.สพ. ดร. ทวีศักดิ์" (Assistant Prof. Dr. Taweesak) and cut off his last name.

Moreover, in Thai, lots of names, especially location and organization names, are in the same forms as general phrases. Consequently, if they do not appear in the obvious context which can indicate that they are actually proper names, the systems could not differentiate between them and general phrases. For example, the system could not recognize "ศูนย์ปฏิบัติการโรคไข้หวัดนก" (Bird Flu Operation Centre) or "องค์การเฝ้าระวังโรคระบาดในสัตว์" (The Animal Disease Surveillance Organization) as the organization names. This explains why the accuracy of person names is higher than those of other names. It is possible to enhance the system performance by using more syntactic features like part of speeches, but for cases in which named entities have the same word forms and syntactic structures like general words, whether those named entity would be recognized is still a question.

Conclusion

In this study, we present Thai named entity recognition system using CRFs as well as comparing the performance of the systems applying different answer patterns. The results show that the answer pattern does have an effect on the performance of the system. The answer patterns having more information can help improve the system performance than those having less information. Nevertheless, the systems have the problem of extracting names without clue words or in ambiguous context. Thus, for future work, we may consider using POS as a feature in our model.

References

Aroonmanakun, W. 2002. Collocation and Thai Word Segmentation. *Proceedings of SNLP-Oriental COCOSDA 2002*. pp.68-75. Prachuapkhirikhan.

Chanlekha, H., and A. Kawtrakul. 2004. Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. *International Joint Conference of Natural Language Processing (IJCNLP-2004).* Hainan Island.

Chanlekha, H., A. Kawtrakul, P. Varasrai and I. Mulasas. 2002. Statistical and Heuristic Rule Based Model for Thai Named Entity Recognition. *Proceeding of SNLP- Oriental COCOSDA 2002*. Hua Hin.

Charoenpornsawat, P., B. Kijsirikul and S. Meknavin. 1998. Feature-based Proper Name Identification in Thai. *Proceedings of National Computer Science and Engineering Conference.* Bangkok.

Feng, Y., L. Sun and Y. Lv. 2006. Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields Models. *Proceeding of the Fifth SIGHAN Workshop on Chinese Language Processing*, pp. 181-184. Sydney.

Haruechaiyasak, C., S. Kongyoung and M. N. Dailey. 2008. A Comparative Study on Thai Word Segmentation Approach. *Proceedings of ECTI-CON*. Krabi.

Lafferty, J., A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Sequence Data. *Proceeding of 18thICML*. San Francisco.

Mao, X., S. He, S. Bao, Y. Dong and H. Wang. 2008. Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields. *Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing*, pp. 90-93. Hyderabad.

Sansing, Y., and A. Prayot. 2010. Thai Named Entity Recognition using Hidden Markov Model. *Proceedings of the Second Conference on Knowledge and Smart Technologies*. Chonburi. [in Thai]

Suwanno, N., Y. Suzuki and H. Yamazaki. 2007. Selecting the optimal feature sets for Thai named entity extraction. *Proceedings of ICEE-2007 & PEC-5*. Phuket.

Tepdang, S., C. Haruechaiyasak, and R. Kongkachandra. (n.d.) *Thai Named Entity Recognition by using Conditional Random Fields*. (n.p.)

Tirasaroj, N. and W. Aroonmanakun. 2009. Thai Named Entity Recognition Based on Conditional Random Fields. *Proceedings of the Eighth International Symposium on Natural Language Processing*. Bangkok.

Yu, X., W. Lam, S. Chan, Y. Wu and B. Chen. 2008. Chinese NER Using CRFs and Logic for the Fourth SIGHAN Bakeoff. *Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing*, pp. 102-105. Hyderabad.