Domain Adaptation with Filtering for Named Entity Extraction of Japanese Anime-Related Words

Kanako KOMIYA [1] **Daichi EDAMURA** [2] **Ryuta TAMURA** [2]
Minoru SASAKI [1] **Hiroyuki SHINNOU** [1] **Yoshiyuki KOTANI** [2]
Ibaraki University [1] Tokyo University of Agriculture and Technology [2]
{kkomiya, msasaki, shinnou}@mx.ibaraki.ac.jp
edaedaichi@gmail.com, {50015646125@st, kotani@cc}.tuat.ac.jp

Abstract

We developed a system to extract Japanese anime-related words, i.e., Japanese NEs (named entities) in the anime-related domain. Since the NEs in the area, such as the titles of anime or the names of characters, were domain-specific, we started by building a tagged corpus and then used it for the experiments. We examined to see if the existing corpora were useful to improve the results. The experiments conducted using Conditional Random Fields showed that the effect of domain adaptation varied according to the genres of the corpora, but the filtering of the source data not only reduced the time for training but also assisted in the domain adaptation work.

1 Introduction

Japanese pop culture such as that exemplified by manga, anime, and gaming has recently gained popularity with the younger generations. They are commercially important; if the NEs (named entities) in the area, such as the titles of the anime and the names of the character could be automatically obtained, they would be useful for the product search, identification, or recommendation. Therefore, we developed a system to extract Japanese NEs in these areas using Conditional Random Fields (CRF). Since the NEs in the area (see Section 3), such as the titles of the anime or the names of the characters, are domain-specific, we build a tagged corpus in the anime domain (see Section 5). We examined to see if the existing corpora were useful with the anime corpus using domain adaptation, since tagging of corpora is time-consuming. The experiments (see Sections 4 and 6) showed that the effect of domain adaptation varied according to the genres of the corpora, but filtering the source data not only reduced the time for training but also asisted in the domain adaptation work. The F-measure was the best when the newspaper and anime corpora were used simultaneously with the domain adaptation after the filtering of the newspaper data (see Section 7).

2 Related Work

Named entity recognition (NER), which involves seeking to locate and classify elements in text into predefined categories, such as the names of people, organizations, and locations, has recently been intensively studied. There are two types of NER methods, i.e., NER using pattern matching and NER using supervised machine learning. NER using pattern matching finds elements in text that match the manually predefined patterns, i.e., the character strings that tend to co-occur with the NEs, e.g., “Mr.” or “University” (Takemoto et al., 2001). There are some works on analyzing or creating these patterns; (Lertcheva and Aroonmanakun, 2011) have analyzed the patterns of the product names used in Thai economic news. NER using pattern matching can extract the NEs that precisely match the patterns, but cannot extract the NEs that do not match the patterns. Therefore, it is difficult to use these types of methods in our system because the anime-related NEs such as the titles of an anime often do not match the patterns.

On the other hand, NER using supervised machine learning trains the patterns to extract the NEs using a tagged corpus. (Yamada et al., 2002) have carried out NER using a support vector machine (SVM). (Nakano and Hirai, 2004) have proposed conducting NER using a bunsetsu feature. Considerable achievements have been made using these methods. In addition, the Hidden Markov Model (HMM) and CRF are often used for NER (Ponomareva et al., 2007), (Ekbal and Bandyopadhyay, 2007). (Asahara and Matsumoto, 2004) have also proposed a character-based chunking method to address the unit problem where NER using su-

Proceedings of Recent Advances in Natural Language Processing, pages 291–297,
Hissar, Bulgaria, Sep 7–9 2015.

pervised machine learning in Japanese originally cannot extract NEs that are smaller than morphemes because cascading morphological analysis and chunking is usually used for any NE extraction in Japanese. (An et al., 2003) have automatically collected NE tagged corpora from World Wide Web to alleviate the problem: building corpus is time-consuming.

There are some works on the adaptation of NER. (Shen et al., 2003) have investigated the effective features of a Hidden Markov model-based NE recognizer for the biomedical domain. (Chiticariu et al., 2010) have improved NER rule language (NERL) for the pattern-based domain adaptation of NER. (Guo et al., 2009) have proposed a domain adaptation method using latent semantic association.

We developed a system to extract Japanese anime-related words using machine learning method, i.e., CRF, for this paper. Since our purpose is to use the anime-related words for the product search, identification, or recommendation, we only extracted them and did not automatically classify them into sub-classes. We examined to see if the existing corpora were useful with the corpus that we built using domain adaptation and showed that the filtering of the source data assisted in the domain adaptation work.

3 Definition of Anime-related Words

We defined the anime-related NEs based on Sekine's extended NE hierarchy (Sekine, 2008). The time and numerical representations were removed because they usually do not appear only in anime but also in real life. Place names were also removed because it is difficult to distinguish place names that appear only in anime from those that appear in real life.

We had two kinds of anime-related words: interior and exterior. The former contains the titles of anime and the anime-related NEs that appear in the anime, and the latter is those that do not, such as the animators. Our system covered both of these. The interior anime-related words include the titles of anime, the names of characters, animals, imaginary creatures, gods, organizations, facilities, products, events, natural objects such as stones, and states such as diseases that appear in the anime. The exterior anime-related words include the names of people such as the authors of the original story, animators, and game creators, the names of organizations such as the production companies and broadcasting companies, the names of related products, and the names of relevant sites, related events, and so on. Table 1 lists some examples of the anime-related words.

4 System to Extract Anime-related Words

The CRF was used as a sequential labeling method to extract anime-related words. There are four steps in the extraction of anime-related words:

1. The parameters are learned through the training corpus.
2. When the text is input into the system, the morphological analysis is carried out and the features are automatically generated.
3. Sequential labeling is carried out using CRF based on the generated features.
4. The NEs are extracted with tags.

We used the BIESO tags (B-beginning, I-intermediate, E-end, S-single token concept, O-outside), for the CRF. Five types of feature, i.e., morpheme, Part-of-speech (POS), finer subcategory of POS, character type, and No. of characters were introduced to train the model of the CRF. They were extracted from the surrounding words of the target morpheme. We used the character type as a feature because the Japanese language has many types of characters and it seemed to be related to the ability of the morphemes to be NEs, especially for anime-related words. The values of the types are hiragana, katakana, alphabetical letters, Chinese characters, and others including punctuation marks. We used the type of the initial character of the morpheme for this feature.

5 Data

We used an anime corpus that we built for the experiments. The texts consisted of 50 anime articles from Wikipedia. The morphological analysis was automatically carried out but the errors in the word segmentations and the POS tags of personal names were manually corrected. After that, all the NEs that we defined above were manually annotated. We used the extended NE tagged corpora (Hashimoto et al., 2008), which were based on the Balanced Corpus of Contemporary Japanese (BCCWJ) (Maekawa, 2008), for a reference when we built the corpus.

Detailed conception	Example	Translation or Explanation
Name of animal character	ポチ	Pochi(Dog's name)
Name of special weapon	かめはめ波	Kame Hame Ha
Name of character	宿海仁太	Yadomi Zhinta
Nickname of character	じんたん	Zintan

Table 1: Examples of anime-related words

The anime-related NE tagged BCCWJ were created based on the extended NE tagged corpora on BCCWJ and they were also used for the training data. We investigated to see if they could be used for the training data on their own and could be used with the corpus that we built with and without domain adaptation. Table 2 summarizes the number of characters, morphemes, NEs, and O tags in the anime corpus and the BCCWJ. The number of O tags after filtering, where all the tokens with an O tag outside of the window of the NEs were filtered out, is also itemized in the table. The genres we used in the BCCWJ are Q&A sites, blogs, novels, magazines, and newspapers. Although the POS were manually annotated on the corpora, the morphological analysis was automatically carried out on them when we used them for the training data; the POS tags should be the same as the anime corpus to train the CRF. After that, the morphemes that have NE tags similar to those of the anime corpus were listed. Then, the BIESO tags were automatically annotated on all the morphemes in the BCCWJ, based on their orthography or spelling using the list of NEs. The NE tags that we used were Product_Other, Character, Doctrine_Method_Other, Company, Broadcast_Program, Occasion_Other, Person, Show_Organization, Movie, Company_group, School, Organization_Other, Country, Music, Offense, Book, National_Language, Event_Other, Class, Food_Other, Corporation_Other, Ethnic_Group_Other, Animal_Disease, Period_Other, Award, Clothing, Magazine, Military, and Name_Other.

Although the NEs that are irrelevant to anime are not tagged on the anime corpus, because they are outside the scope of our research, the BCCWJ contains many of them. Therefore, the anime corpus has many NEs whose POS subcategories are proper name whereas the BCCWJ have many general noun ones.

6 Experiment

CRF++ [1] and MeCab [2] were used as the CRF tool and as a morphological analyzer, respectively. The morphemes were used as tokens in CRF++ and each of the alphabetical words was processed as one token. The parameters *f* and *c* in CRF++ were set to two and one, respectively, according to the results from preliminary experiments.

We used three types of features, i.e., the morphemes, the POS, and the finer subcategory of POS inside a window size of 5, the character type inside a window size of 3, and the number of the characters inside a window size of 1. These window sizes were determined according to the results from preliminary experiments.

Table 3 specifies an example of the tagged anime corpus. The meanings of the morphemes are also shown in the table for reference. If "戦い" (Fight) was the target morpheme, the features within "帝国" (Empire), "と" (Against), "戦い" (Fight), "、" (,). and "未知" (Unknown) are used for the three types of features, the features within "と" (Against), "戦い" (Fight) and, "、" (,) are used for the character type, and the features within "戦い" (Fight) are used for the number of the characters.

We used input-level unigrams, bigrams, trigrams, and 4grams for the POS and the finer subcategories of POS within the window size of 5. However, only the unigrams in the window size of 5 and the bigrams in the window size of 3 were used for the morphemes, because their combination number would be extremely large. The combination of the POS and the finer subcategory of POS of the target morpheme, and the combinations of the previous output and target morpheme were also used in all the experiments. When the character type and the number of characters were used, POS and morpheme combination, and that of the finer subcategory of POS and the morpheme

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html
[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

Type	Anime	Q&A site	Blog	Novel	Magazine	Newspaper
Total No. of characters	44,829	177,636	189,474	369,345	388,941	56,2206
Total No. of morphemes	26,948	108,182	116,885	228,721	236,369	353,882
Total No. of NEs	1,570	2,202	4,173	5,042	7,758	13,629
NEs of S tag	742	1,282	2,772	3,163	4,409	6,928
NEs of BIE tags	828	920	1,401	1,879	3,349	6,701
O tags	23,824	104,377	109,922	220,753	222,259	327,939
O tags after filtering	No filtering	7,792	14,013	18,228	27,080	47,324

Table 2: No. of characters, morphemes, NEs, and O tags in BCCWJ and anime corpus

of the target morpheme were also used. Five-fold cross validation [3] was used in the experiments.

We examined to see if the existing corpora, the BCCWJ, were useful for extracting anime-related words, and if they were useful for the training data on their own and together with the anime corpus, using domain adaptation. The experiments were carried out without domain adaptation, as a reference. (Daumé III, 2007) was used as the domain adaptation method, where an input space was augmented and general, source-specific, and target-specific triple length features were made. The mappings were $\Phi_s(x) =< x, x, 0 >$, $\Phi_t(x) =< x, 0, x >$, where $\Phi_s(x)$ and $\Phi_t(x)$ denoted the mappings to map the source and target data, respectively, $< x > \in \mathbb{R}^F$ was the original input, and $0 =< 0, 0, ..., 0 > \in \mathbb{R}^F$ was the zero vector.

We also investigated to see if the filtering of the source data, where all the tokens with an O tag outside of the window of the NEs were filtered out, assisted in the domain adaptation work. We assumed the recall would be improved when many tokens with O tags were filtered out.

7 Results

The experimental results, i.e., the tag accuracy, the recall, the precision, and the F-measure, according to the corpora and the filtering on a single corpus are listed in Table 4. The experimental results according to the corpora and the filtering when the anime corpus and BCCWJ were used together with a simple augmentation and using Daumé's method of domain adaptation, respectively, are summarized in Tables 5 and 6. The results in bold denote the results that were superior to those of the system trained using only the anime corpus and the underline means the value is the best result overall.

The recall, precision, and F-measure were evaluated based on chunks. In other words, the NEs with a BIE tag are correct only if all the tags from the initial B to the last E in the chunk are correct. The tag accuracy is the number of correct tags in the total number of tags.

8 Discussion

First, let us focus on the results with no filtering. The results listed in Table 4 show that the recalls are very low and the precisions are not very good when the BCCWJ are used for the training on their own. According to Table 5, when the BCCWJ and anime corpus are used together using simple augmentation, the recalls are not very good but the precisions are comparable to (The average is slightly better than) the results when using only the anime corpus. Finally, the results in Table 6 show that the recalls are slightly better than and the precisions are slightly worse than the results when only the anime corpus is used. However, the averaged F-measure is slightly worse than that of the anime corpus. These results show that the domain adaptation slightly improved the recalls and reduced the level of precision and that the F-measures did not change very much.

Next, let us consider the experimental results when using the filtering. The results in Table 4 show that the recalls greatly improved but the precisions considerably decreased when the filtering was used. We think this is because many tokens with O tags were deleted; it makes the system extract more NEs. We can see from Table 5 that the situation is almost the same, when the corpora are used together with the simple augmentation: the recalls improved but the precision decreased when the filtering was used, which has no effect on the F-measures. According to Table 6, the improvement of the recalls is not very large

[3]The full source corpus and the four fifth anime corpus were used for the training and the rest one fifth anime corpus was used for the test.

Meaning	Morpheme	POS	Finer subcategory of POS	Char. type	N of Chars	Tag
Yamato	ヤマト	Noun	Proper name-organization	Katakana	3	S
Topic-marking	は	Particle	Linking particle	Hiragana	1	o
Gamirasu	ガミラス	Noun	General	Katakana	4	B
Empire	帝国	Noun	General	Chinese	2	E
Against	と	Particle	Case-marking-general	Hiragana	1	o
Fight	戦い	Verb	Independent word	Chinese	2	o
,	、	Mark	Punctuation	Punctuation	1	o
Unknown	未知	Noun	General	Chinese	2	o
Of	の	Particle	Adnominalize	Hiragana	1	o
Universe	宇宙	Noun	General	Chinese	2	o
Space	空間	Noun	General	Chinese	2	o
In	における	Particle	Case-marking-collocation	Hiragana	4	o
Obstacle	障害	Noun	General	Chinese	2	o
Object-marking	を	Particle	Case-marking-general	Hiragana	1	o
Overcome	乗り越え	Verb	Independent word	Chinese	4	o

Table 3: Example of tagged anime corpus

Filtering	Corpora	Tag accuracy	Recall	Precision	F-measure
No	Anime	94.92%	68.47%	84.65%	75.70%
No	Q&A site	91.59%	33.95%	70.88%	45.91%
No	Blog	92.19%	42.80%	77.42%	55.13%
No	Novel	93.16%	48.28%	82.93%	61.03%
No	Magazine	93.50%	48.47%	**86.18%**	62.05%
No	Newspaper	92.64%	46.31%	76.69%	57.74%
No	Avg.	92.62%	43.96%	78.82%	56.37%
Yes	Q&A site	78.76%	**72.04%**	26.57%	38.82%
Yes	Blog	81.36%	**77.13%**	30.95%	44.17%
Yes	Novel	93.07%	**80.45%**	30.20%	60.45%
Yes	Magazine	83.19%	**80.83%**	32.29%	46.15%
Yes	Newspaper	81.76%	**76.11%**	30.23%	43.27%
Yes	Avg.	81.36%	**77.31%**	30.05%	43.27%

Table 4: Experimental results according to corpora and filtering on single corpus

Filtering	Corpora	Tag accuracy	Recall	Precision	F-measure
No	Q&A	94.55%	62.80%	83.56%	71.71%
No	Blog	94.40%	61.85%	84.14%	71.29%
No	Novel	94.36%	61.15%	**85.18%**	71.19%
No	Magazine	94.62%	60.06%	**88.13%**	71.44%
No	Newspaper	94.24%	58.79%	83.08%	68.85%
No	Avg.	94.43%	60.93%	**84.82%**	70.90%
Yes	Q&A	94.21%	**75.10%**	70.85%	72.91%
Yes	Blog	94.70%	**76.43%**	73.22%	74.79%
Yes	Novel	94.33%	**77.96%**	68.88%	73.14%
Yes	Magazine	94.38%	**79.68%**	69.38%	74.18%
Yes	Newspaper	93.75%	**78.09%**	66.20%	71.65%
Yes	Avg.	94.28%	**77.45%**	69.71%	73.33%

Table 5: Experimental results according to corpora and filtering using simple augmentation

Filtering	Corpora	Tag accuracy	Recall	Precision	F-measure
No	Q&A	**94.99%**	**69.17%**	83.73%	**75.76%**
No	Blog	**94.95%**	**69.36%**	83.26%	75.68%
No	Novel	**94.95%**	**68.98%**	83.63%	75.60%
No	Magazine	94.80%	68.09%	83.19%	74.89%
No	Newspaper	94.89%	**69.11%**	83.59%	75.66%
No	Avg.	94.92%	**68.94%**	83.48%	75.52%
Yes	Q&A	**94.95%**	**69.30%**	83.95%	**75.92%**
Yes	Blog	**95.01%**	**68.92%**	83.42%	75.48%
Yes	Novel	**95.00%**	**69.55%**	83.94%	**76.07%**
Yes	Magazine	**95.11%**	**69.62%**	84.53%	**76.35%**
Yes	Newspaper	**95.08%**	**70.13%**	83.92%	**76.41%**
Yes	Avg.	**95.03%**	**69.50%**	83.95%	**76.05%**

Table 6: Experimental results according to corpora and filtering using domain adaptation

but the decrease in the level of precision is also not very large, because the degree of improvement increased and the degree of decrease lessened when using the filtering. However, the F-measures improved. These results show that the filtering with domain adaptation could improve the recalls while not affecting the level of precision too much.

Method	Filtering	S	BIE
Original	No	436.8	436.2
Original	Yes	2,279.0	1,768.0
Simple aug.	No	589.6	538.8
Simple aug.	Yes	883.4	863.6
DA	No	658.0	638.6
DA	Yes	663.6	636.2

Table 7: Averaged number of NEs that system output

As described above, the filtering made the system extract more NEs. Table 7 lists the averaged number of the NEs that the system extracted. The filtering did not affect the number of NEs that the system output when domain adaptation was used, but the numbers of correct answers increased by the filtering.

The results in Tables 4, 5, and 6 show that only the systems using domain adaptation can outperform the system trained using only the anime corpus. In addition, the results in Table 6 show that the effect of the domain adaptation varies according to the genre of the corpora; only the Q&A site data could improve the F-measure of the system without filtering. However, the other results in this table show that four-fifths of the system improved the F-measures. The F-measure was the best when the newspaper and anime corpora were used together using the domain adaptation after the newspaper data was filtered.

Finally, the filtering has another advantage: the time for training, which was reduced to only 15% of that of the system trained with full corpora.

9 Conclusion

We developed a system to extract Japanese anime-related words using CRF and examined to see if the corpora whose genre were not anime were useful for improving the results. We investigated to see if they could be used for the training data on their own and could be used with the anime corpus that we built with and without domain adaptation. We also examined to see if the filtering of the source data, where all the tokens with an O tag outside of the window of the NEs were filtered out, assisted the domain adaptation work. The experiments showed that (1) the non-anime corpora could improve the F-measure when they were used with the anime corpus using only domain adaptation, (2) the effect of the domain adaptation varies according to the genre of the corpora, and (3) the domain adaptation with the filtering improved the recalls without effecting the level of precision too much, which improved the F-measure. Moreover, the training time was reduced to only 15% of that of the system trained with full corpora.

Acknowledgment

We thank Mr. Masaki TAKASE, who built the annime corpus and developed a baseline system. This work was partially supported by JSPS KAKENHI Grant Number 24700138.

References

Joohui An, Seungwoo Lee, and Gary Geunbae Lee. 2003. Automatic acquisition of named entity tagged corpus from world wide web. In *Proc. of ACL'03*, pages 165–168.

Masayuki Asahara and Yuji Matsumoto. 2004. A word unit problem in japanese named entity extraction. *IPSJ Journal (In Japanese)*, Vol. 45(No. 5):1442–1450.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on EMNLP*, pages 1002–1012.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Asif Ekbal and Sivaji Bandyopadhyay. 2007. A hidden markov model based named entity recognition system: Bengali and hindi as case studies. In *Proceedings of the 2nd international conference on Pattern recognition and machine intelligence*, pages 545–552.

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Xian Wu, and Zhong Su. 2009. Domain adaptation with latent semantic association for named entity recognition. In *Proceedings of the 2009 Annual Conference of the NAACL*, pages 281–289.

Taiichi Hashimoto, Takashi Inui, and Koji Murakami. 2008. Constructing extended named entity annotated corpora. In *IPSJ SIG Notes 2008 (In Japanese)*, pages 113–120.

Nattadaporn Lertcheva and Wirote Aroonmanakun. 2011. Product name identification and classification in thai economic news. In *Proc. of IJCNLP 2011 Named Entities Workshop*, pages 58–61.

Kikuo Maekawa. 2008. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pages 101–102.

Keigo Nakano and Yuzo Hirai. 2004. Japanese named entity extraction with bunsetsu features. *IPSJ Journal (In Japanese)*, Vol. 45(No. 3):934–941.

Natalia Ponomareva, Paolo Rosso, Ferran Pla, and Antonio Molina. 2007. Conditional random fields vs. hidden markov models in a biomedical named entity recognition task. In *Proceedings of the Recent Advances in Natural Language Processing 2007*, pages 1–7.

Satoshi Sekine. 2008. Extended named entity ontology with attribute information. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 52–57.

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2003. Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. In *Proceedings of the Whorkshop on NLP Bionmedicine, ACL*, pages 49–56.

Yoshikazu Takemoto, Toshikazu Fukushima, and Hiroshi Yamada. 2001. A japanese named entity extraction system based on building a large-scale and high-quality dictionary and pattern-matching rules. *IPSJ Journal (In Japanese)*, Vol. 42(No. 6):1580–1591.

Hiroyasu Yamada, Taku Kudo, and Yuji Matsumoto. 2002. Japanese named entity extraction using support vector machine. *IPSJ Journal (In Japanese)*, Vol. 43(No. 1):44–53.