Automatic Acquisition of Named Entity Tagged Corpus from World Wide Web

Joohui An
Dept. of CSE
POSTECH
Pohang, Korea 790-784
minnie@postech.ac.kr

Seungwoo Lee
Dept. of CSE
POSTECH
Pohang, Korea 790-784
pinesnow@postech.ac.kr

Gary Geunbae Lee
Dept. of CSE
POSTECH
Pohang, Korea 790-784
gblee@postech.ac.kr

Abstract

In this paper, we present a method that automatically constructs a Named Entity (NE) tagged corpus from the web to be used for learning of Named Entity Recognition systems. We use an NE list and an web search engine to collect web documents which contain the NE instances. The documents are refined through sentence separation and text refinement procedures and NE instances are finally tagged with the appropriate NE categories. Our experiments demonstrates that the suggested method can acquire enough NE tagged corpus equally useful to the manually tagged one without any human intervention.

1 Introduction

Current trend in Named Entity Recognition (NER) is to apply machine learning approach, which is more attractive because it is trainable and adaptable, and subsequently the porting of a machine learning system to another domain is much easier than that of a rule-based one. Various supervised learning methods for Named Entity (NE) tasks were successfully applied and have shown reasonably satisfiable performance.((Zhou and Su, 2002)(Borthwick et al., 1998)(Sassano and Utsuro, 2000)) However, most of these systems heavily rely on a tagged corpus for training. For a machine learning approach, a large corpus is required to circumvent the data sparseness problem, but the dilemma is that the costs required to annotate a large training corpus are non-trivial.

In this paper, we suggest a method that automatically constructs an NE tagged corpus from the web to be used for learning of NER systems. We use an NE list and an web search engine to collect web documents which contain the NE instances. The documents are refined through the sentence separation and text refinement procedures and NE instances are finally annotated with the appropriate NE categories. This automatically tagged corpus may have lower quality than the manually tagged ones but its size can be almost infinitely increased without any human efforts. To verify the usefulness of the constructed NE tagged corpus, we apply it to a learning of NER system and compare the results with the manually tagged corpus.

2 Automatic Acquisition of an NE Tagged Corpus

We only focus on the three major NE categories (i.e., person, organization and location) because others are relatively easier to recognize and these three categories actually suffer from the shortage of an NE tagged corpus.

Various linguistic information is already held in common in written form on the web and its quantity is recently increasing to an almost unlimited extent. The web can be regarded as an infinite language resource which contains various NE instances with diverse contexts. It is the key idea that automatically marks such NE instances with appropriate category labels using pre-compiled NE lists. However, there should be some general and language-specific con-

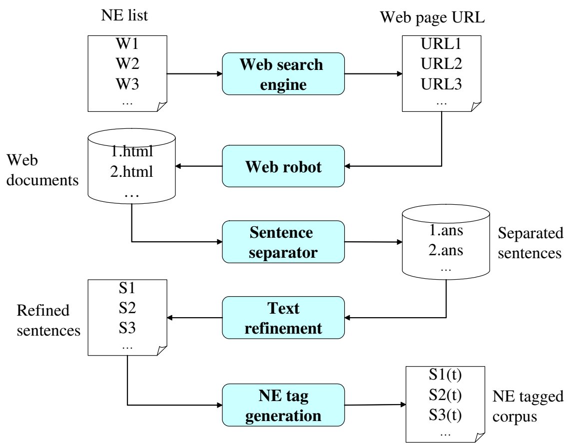

Figure 1: Automatic generation of NE tagged corpus from the web

siderations in this marking process because of the word ambiguity and boundary ambiguity of NE instances. To overcome these ambiguities, the automatic generation process of NE tagged corpus consists of four steps. The process first collects web documents using a web search engine fed with the NE entries and secondly segments them into sentences. Next, each sentence is refined and filtered out by several heuristics. An NE instance in each sentence is finally tagged with an appropriate NE category label. Figure 1 explains the entire procedure to automatically generate NE tagged corpus.

2.1 Collecting Web Documents

It is not appropriate for our purpose to randomly collect documents from the web. This is because not all web documents actually contain some NE instances and we also do not have the list of all NE instances occurring in the web documents. We need to collect the web documents which necessarily contain at least one NE instance and also should know its category to automatically annotate it. This can be accomplished by using a web search engine queried with pre-compiled NE list.

As queries to a search engine, we used the list of Korean Named Entities composed of 937 person names, 1,000 locations and 1,050 organizations. Using a Part-of-Speech dictionary, we removed ambiguous entries which are not proper nouns in other contexts to reduce errors of automatic annotation. For example, '경기(kyunggi, *Kyunggi/business conditions/a game*)' is filtered out because it means a location (proper noun) in one context, but also means business conditions or a game (common noun) in other contexts. By submitting the NE entries as queries to a search engine[1], we obtained the maximum 500 of URL's for each entry. Then, a web robot visits the web sites in the URL list and fetches the corresponding web documents.

2.2 Splitting into Sentences

Features used in the most NER systems can be classified into two groups according to the distance from a target NE instance. The one includes internal features of NE itself and context features within a small word window or sentence boundary and the other includes name alias and co-reference information beyond a sentence boundary. In fact, it is not easy to extract name alias and co-reference information directly from manually tagged NE corpus and needs additional knowledge or resources. This leads us to focus on automatic annotation in sentence level, not document level. Therefore, in this step, we split the texts of the collected documents into sentences by (Shim et al., 2002) and remove sentences without target NE instances.

2.3 Refining the Web Texts

The collected web documents may include texts actually matched by mistake, because most web search engines for Korean use n-gram, especially, bi-gram matching. This leads us to refine the sentences to exclude these erroneous matches. Sentence refinement is accomplished by three different processes: separation of functional words, segmentation of compound nouns, and verification of the usefulness of the extracted sentences.

An NE is often concatenated with more than one `josa`, a Korean functional word, to compose a Korean word. Therefore we need to separate the functional words from an NE instance to detect the boundary of the NE instance and this is achieved by a part-of-speech tagger, POSTAG, which can detect unknown words (Lee et al., 2002). The separation of functional words gives us another benefit that we can resolve the ambiguities between an NE and a common noun plus functional words

[1] We used Empas (http://www.empas.com)

		Person	Location	Organization
Training	Automatic	29,042	37,480	2,271
	Manual	1,014	724	1,338
Test	Manual	102	72	193

Table 1: Corpus description (number of NE's) (Automatic: Automatically annotated corpus, Manual: Manually annotated corpus

and filter out erroneous matches. For example, '경기도(kyunggi-do)' can be interpreted as either '경기도(*Kyunggi Province*)' or '경기+도(*a game also*)' according to its context. We can remove the sentence containing the latter case.

A josa-separated Korean word can be a compound noun which only contains a target NE as a substring. This requires us to segment the compound noun into several correct single nouns to match with the target NE. If the segmented single nouns are not matched with a target NE, the sentence can be filtered out. For example, we try to search for an NE entry, '핑클(*Fin.KL*, a Korean singer group)' and may actually retrieve sentences including '써핑클럽(*surfing club*)'. The compound noun, '써핑클럽', can be divided into '써핑(*surfing*)' and '클럽(*club*)' by a compound-noun segmenting method (Yun et al., 1997). Since both '써핑' and '클럽' are not matched with our target NE, '핑클', we can delete the sentences. Although a sentence has a correct target NE, if it does not have context information, it is not useful as an NE tagged corpus. We also removed such sentences.

2.4 Generating an NE tagged corpus

The sentences selected by the refining process explained in previous section are finally annotated with the NE label. We acquired the NE tagged corpus including 68,793 NE instances through this automatic annotation process. We can annotate only one NE instance per sentence but almost infinitely increase the size of the corpus because the web provides unlimited data and our process is fully automatic.

3 Experimental Results

3.1 Usefulness of the Automatically Tagged Corpus

For effectiveness of the learning, both the size and the accuracy of the training corpus are important.

Training corpus	Precision	Recall	F-measure
Seeds only	84.13	42.91	63.52
Manual	80.21	86.11	83.16
Automatic	81.45	85.41	83.43
Manual + Automatic	82.03	85.94	83.99

Table 2: Performance of the decision list learning

Generally, the accuracy of automatically created NE tagged corpus is worse than that of hand-made corpus. Therefore, it is important to examine the usefulness of our automatically tagged corpus compared to the manual corpus. We separately trained the decision list learning features using the automatically annotated corpus and hand-made one, and compared the performances. Table 1 shows the details of the corpus used in our experiments.[2]

Through the results in Table 2, we can verify that the performance with the automatic corpus is superior to that with only the seeds and comparable to that with the manual corpus.Moreover, the domain of the manual training corpus is same with that of the test corpus, i.e., news and novels, while the domain of the automatic corpus is unlimited as in the web. This indicates that the performance with the automatic corpus should be regarded as much higher than that with the manual corpus because the performance generally gets worse when we apply the learned system to different domains from the trained ones. Also, the automatic corpus is pretty much self-contained since the performance does not gain much though we use both the manual corpus and the automatic corpus for training.

3.2 Size of the Automatically Tagged Corpus

As another experiment, we tried to investigate how large automatic corpus we should generate to get the satisfiable performance. We measured the performance according to the size of the automatic corpus. We carried out the experiment with the decision list learning method and the result is shown in Table 3. Here, 5% actually corresponds to the size of the manual corpus. When we trained with that size of the automatic corpus, the performance was very low compared to the performance of the manual corpus. The reason is that the automatic corpus is com-

[2]We used the manual corpus used in Seon et al. (2001) as training and test data.

Corpus size (words)	Precision	Recall	F-measure
90,000 (5%)	72.43	6.94	39.69
448,000 (25%)	73.17	41.66	57.42
902,000 (50%)	75.32	61.53	68.43
1,370,000 (75%)	78.23	77.19	77.71
1,800,000 (100%)	81.45	85.41	83.43

Table 3: Performance according to the corpus size

Corpus size (words)	Precision	Recall	F-measure
700,000	79.41	81.82	80.62
1,000,000	82.86	85.29	84.08
1,200,000	83.81	86.27	85.04
1,300,000	83.81	86.27	85.04

Table 4: Saturation point of the performance for 'person' category

posed of the sentences searched with fewer named entities and therefore has less lexical and contextual information than the same size of the manual corpus. However, the automatic generation has a big merit that the size of the corpus can be increased almost infinitely without much cost. From Table 3, we can see that the performance is improved as the size of the automatic corpus gets increased. As a result, the NER system trained with the whole automatic corpus outperforms the NER system trained with the manual corpus.

We also conducted an experiment to examine the saturation point of the performance according to the size of the automatic corpus. This experiment was focused on only 'person' category and the result is shown in Table 4. In the case of 'person' category, we can see that the performance does not increase any more when the corpus size exceeds 1.2 million words.

4 Conclusions

In this paper, we presented a method that automatically generates an NE tagged corpus using enormous web documents. We use an internet search engine with an NE list to collect web documents which may contain the NE instances. The web documents are segmented into sentences and refined through sentence separation and text refinement procedures. The sentences are finally tagged with the NE categories. We experimentally demonstrated that the suggested method could acquire enough NE tagged corpus equally useful to the manual corpus without any human intervention. In the future, we plan to apply more sophisticated natural language processing schemes for automatic generation of more accurate NE tagged corpus.

Acknowledgements

This research was supported by BK21 program of Korea Ministry of Education and MOCIE strategic mid-term funding through ITEP.

References

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 152–160, New Brunswick, New Jersey. Association for Computational Linguistics.

Gary Geunbae Lee, Jeongwon Cha, and Jong-Hyeok Lee. 2002. Syllable Pattern-based Unknown Morpheme Segmentation and Estimation for Hybrid Part-Of-Speech Tagging of Korean. *Computational Linguistics*, 28(1):53–70.

Manabu Sassano and Takehito Utsuro. 2000. Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 705–711, Germany.

Choong-Nyoung Seon, Youngjoong Ko, Jeong-Seok Kim, and Jungyun Seo. 2001. Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pages 229–236, Tokyo, Japan.

Junhyeok Shim, Dongseok Kim, Jeongwon Cha, Gary Geunbae Lee, and Jungyun Seo. 2002. Multi-strategic Integrated Web Document Pre-processing for Sentence and Word Boundary Detection. *Information Processing and Management*, 38(4):509–527.

Bo-Hyun Yun, Min-Jeung Cho, and Hae-Chang Rim. 1997. Segmenting Korean Compound Nouns using Statistical Information and a Preference Rule. *Journal of Korean Information Science Society*, 24(8):900–909.

GuoDong Zhou and Jian Su. 2002. Named Entity Recognition using an HMM-based Chunk Tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 473–480, Philadelphia, USA.