Named Entity Discovery Using Comparable News Articles

Yusuke SHINYAMA and **Satoshi SEKINE**
Computer Science Department
New York University
715, Broadway, 7th Floor
New York, NY, 10003
yusuke@cs.nyu.edu, sekine@cs.nyu.edu

Abstract

In this paper we describe a way to discover Named Entities by using the distribution of words in news articles. Named Entity recognition is an important task for today's natural language applications, but it still suffers from data sparseness. We used an observation that a Named Entity is likely to appear synchronously in several news articles, whereas a common noun is less likely. Exploiting this characteristic, we successfully obtained rare Named Entities with 90% accuracy just by comparing time series distributions of a word in two newspapers. Although the achieved recall is not sufficient yet, we believe that this method can be used to strengthen the lexical knowledge of a Named Entity tagger.

1 Introduction

Recently, Named Entity (NE) recognition has been getting more attention as a basic building block for practical natural language applications. A Named Entity tagger identifies proper expressions such as names, locations and dates in sentences. We are trying to extend this to an Extended Named Entity tagger, which additionally identifies some common nouns such as disease names or products. We believe that identifying these names is useful for many applications such as information extraction or question answering (Sekine et al., 2002).

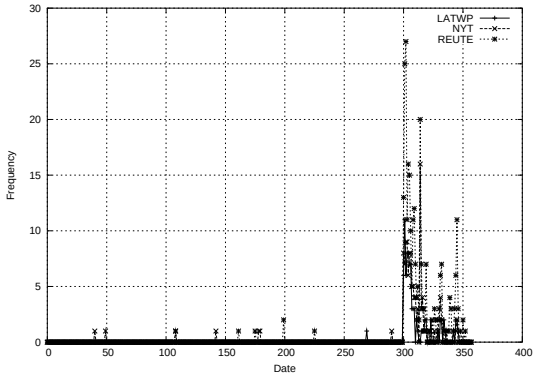
Normally a Named Entity tagger uses lexical or contextual knowledge to spot names which appear in documents. One of the major problem of this task is its data sparseness. Names appear very frequently in regularly updated documents such as news articles or web pages. They are, however, much more varied than common nouns, and changing continuously. Since it is hard to construct a set of predefined names by hand, usually some corpus based approaches are used for building such taggers.

However, as Zipf's law indicates, most of the names which occupy a large portion of vocabulary are rarely used. So it is hard for Named Entity tagger developers to keep up with a contemporary set of words, even though a large number of documents are provided for learning. There still might be a "wild" noun which doesn't appear in the corpora. Several attempts have been made to tackle this problem by using unsupervised learning techniques, which make vast amount of corpora available to use. (Strzalkowski and Wang, 1996) and (Collins and Singer, 1999) tried to obtain either lexical or contextual knowledge from a seed given by hand. They trained the two different kind of knowledge alternately at each iteration of training. (Yangarber et al., 2002) tried to discover names with a similar method. However, these methods still suffer in the situation where the number of occurrences of a certain name is rather small.

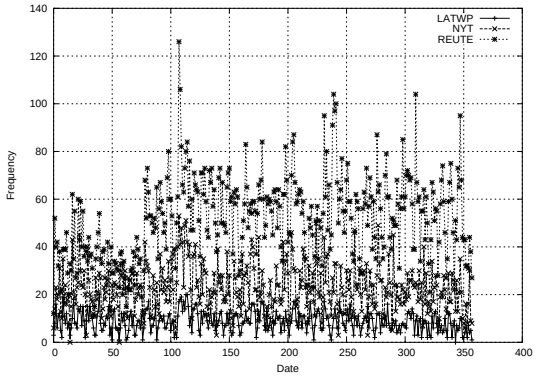
2 Synchronicity of Names

In this paper we propose another method to strengthen the lexical knowledge for Named Entity tagging by using synchronicity of names in comparable documents. One can view a "comparable" document as an alternative expression of the same content. Now, two document sets where each document of one set is associated with one in the other set, is called a "comparable" corpus. A comparable corpus is less restricted than a parallel corpus and usually more available. Several different newspapers published on the same day report lots of the same events, therefore contain a number of comparable documents. One can also take another view of a comparable corpus, which is a set of paraphrased documents. By exploiting this feature, one can extract paraphrastic expressions automatically from parallel corpora (Barzilay and McKeown, 2001) or comparable corpora (Shinyama and Sekine, 2003).

Named Entities in comparable documents have one notable characteristic: they tend to be preserved across comparable documents because it is generally difficult to paraphrase names. We think that it is also hard to paraphrase product names or disease names, so they will also be preserved. Therefore, if one Named Entity appears in one document, it should also appear in the comparable document.



The occurrence of the word "yigal"



The occurrence of the word "killed"

Figure 1: The occurrence of two words in 1995

Consequently, if one has two sets of documents which are associated with each other, the distribution of a certain name in one document set should look similar to the distribution of the name in the other document set.

We tried to use this characteristic of Named Entities to discover rare names from comparable news articles. We particularly focused on the time series distribution of a certain word in two newspapers. We hypothesized that if a Named Entity is used in two newspapers, it should appear in both newspapers synchronously, whereas other words don't. Since news articles are divided day by day, it is easy to obtain the time series distribution of words appearing in each newspaper.

Figure 2 shows the time series distribution of the two words "yigal" and "killed", which appeared in several newspapers in 1995. The word "yigal" (the name of the man who killed Israeli Prime Minister Yitzhak Rabin on Nov. 7, 1995) has a clear spike. There were a total of 363 documents which included the word that year and its occurrence is synchronous between the two newspapers. In contrast, the word "killed", which appeared in 21591 documents, is spread over all the year and has no clear characteristic.

3 Experiment

To verify our hypothesis, we conducted an experiment to measure the correlation between the occurrence of Named Entity and its similarity of time series distribution between two newspapers.

First, we picked a rare word, then obtained its document frequency which is the number of articles which contain the word. Since newspaper articles are provided separately day by day, we sampled the document frequency for each day. These numbers form, for one year for example, a 365-element integer vector per newspaper. The actual number of news articles is oscillating weekly, however, we normalized this by dividing the number of articles containing the word by the total number of all articles on that day. At the end we get a vector of fractions which range from 0.0 to 1.0.

Next we compared these vectors and calculated the similarity of their time series distributions across different news sources. Our basic strategy was to use the cosine similarity of two vectors as the likelihood of the word's being a Named Entity. However, several issues arose in trying to apply this directly. Firstly, it is not always true that the same event is reported on the same day. An actual newspaper sometimes has a one or two-day time lag depending on the news. To alleviate this effect, we applied a simple smoothing to each vector. Secondly, we needed to focus on the salient use of each word, otherwise a common noun which constantly appears almost every day has an undesirable high similarity between newspapers. To avoid this, we tried to intensify the effect of a spike by comparing the deviation of the frequency instead of the frequency itself. This way we can degrade the similarity of a word which has a "flat" distribution.

In this section we first explain a single-word experiment which detects Named Entities that consist of one word. Next we explain a multi-word experiment which detects Named Entities that consist of exactly two words.

3.1 Single-word Experiment

In a single-word experiment, we used two one-year newspapers, Los Angeles Times and Reuters in 1995. First we picked a rare word which appeared in either newspaper less than 100 times throughout the year. We only used a simple tokenizer and converted all words into lower case. A part of speech tagger was not used. Then we obtained the document frequency vector for the word. For each word w which appeared in newspaper A, we got the document frequency at date t:

$$f_A(w,t) = df_A(w,t)/N_A(t)$$

where $df_A(w,t)$ is the number of documents which contain the word w at date t in newspaper A. The normalization constant $N_A(t)$ is the number of all articles at date t. However comparing this value between two newspapers directly cannot capture a time lag. So now we apply smoothing by the following formula to get an improved version of f_A:

$$f'_A(w,t) = \sum_{-W \leq i \leq W} r^{|i|} f_A(w, t+i)$$

Here we give each occurrence of a word a "stretch" which sustains for W days. This way we can capture two occurrences which appear on slightly different days. In this experiment, we used $W = 2$ and $r = 0.3$, which sums up the numbers in a 5-day window. It gives each occurrence a 5-day stretch which is exponentially decreasing.

Then we make another modification to f'_A by computing the deviation of f'_A to intensify a spike:

$$f''_A(w,t) = \frac{f'_A(w,t) - \bar{f'_A}}{\sigma}$$

where $\bar{f'_A}$ and σ is the average and the standard deviation of $f'_A(w)$:

$$\bar{f'_A} = \frac{\sum_t f'_A(w,t)}{T}$$

$$\sigma = \sqrt{\frac{\sum_t (f'_A(w,t) - \bar{f'_A})^2}{T}}$$

T is the number of days used in the experiment, e.g. $T = 365$ for one year. Now we have a time series vector $F_A(w)$ for word w in newspaper A:

$$F_A(w) = \{f''_A(w,1), f''_A(w,2), ..., f''_A(w,T)\}$$

Similarly, we calculated another time series $F_B(w)$ for newspaper B. Finally we computed $sim(w)$, the cosine similarity of two distributions of the word w with the following formula:

$$sim(w) = \frac{F_A(w) \cdot F_B(w)}{|F_A(w)||F_B(w)|}$$

Since this is the cosine of the angle formed by the two vectors, the obtained similarity ranges from -1.0 to 1.0. We used $sim(w)$ as the Named Entity score of the word and ranked these words by this score. Then we took the highly ranked words as Named Entities.

3.2 Multi-word Experiment

We also tried a similar experiment for compound words. To avoid chunking errors, we picked all consecutive two-word pairs which appeared in both newspapers, without using any part of speech tagger or chunker. Word pairs which include a pre-defined stop word such as "the" or "with" were eliminated. As with the single-word experiment, we measured the similarity between the time series distributions for a word pair in two newspapers. One different point is that we compared three newspapers [1] rather than two, to gain more accuracy. Now the ranking score $sim(w)$ given to a word pair is calculated as follows:

$$sim(w) = sim_{AB}(w) \times sim_{BC}(w) \times sim_{AC}(w)$$

where $sim_{XY}(w)$ is the similarity of the distributions between two newspapers X and Y, which can be computed with the formula used in the single-word experiment. To avoid incorrectly multiplying two negative similarities, a negative similarity is treated as zero.

4 Evaluation and Discussion

To evaluate the performance, we ranked 966 single words and 810 consecutive word pairs which are randomly selected. We measured how many Named Entities are included in the highly ranked words. We manually classified as names the words in the following categories used in IREX (Sekine and Isahara, 2000): PERSON, ORGANIZATION, LOCATION, and PRODUCT. In both experiments, we regarded a name which can stand itself as a correct Named Entity, even if it doesn't stretch to the entire noun phrase.

4.1 Single-word Experiment

Table 1 shows an excerpt of the ranking result. For each word, the type of the word, the document frequency and the similarity (score) $sim(w)$ is listed. Obvious typos are classified as "typo". One can observe that a word which is highly ranked is more likely a Named Entity than lower ones. To show this correlation clearly, we plot the score of the words and the likelihood of being a Named Entity in Figure 2. Since the actual number of the words is discrete, we computed the likelihood by counting Named Entities in a 50-word window around that score.

Table 3 shows the number of obtained Named Entities. By taking highly ranked words ($sim(w) \geq$

[1] For the multi-word experiment, we used Los Angeles Times, Reuters, and New York Times.

Word	Type	Freq.	Score
sykesville	LOCATION	4	1.000
khamad	PERSON	4	1.000
zhitarenko	PERSON	6	1.000
sirica	PERSON	9	1.000
energiyas	PRODUCT	4	1.000
hulya	PERSON	5	1.000
salvis	PERSON	5	0.960
geagea	PERSON	27	0.956
bogdanor	PERSON	6	0.944
gomilevsky	PERSON	6	0.939
kulcsar	PERSON	15	0.926
carseats	noun	17	0.912
wilsons	PERSON	32	0.897
yeud	ORGANIZATION	10	0.893
yigal	PERSON	490	0.878
bushey	PERSON	10	0.874
pardew	PERSON	17	0.857
yids	PERSON	5	0.844
bordon	PERSON	113	0.822
...	...	...	...
katyushas	PRODUCT	56	0.516
solzhenitsyn	PERSON	81	0.490
scheuer	PERSON	9	0.478
morgue	noun	340	0.456
mudslides	noun	151	0.420
rump	noun	642	0.417
grandstands	noun	42	0.407
overslept	verb	51	0.401
lehrmann	PERSON	13	0.391
...	...	...	...
willowby	PERSON	3	0.000
unknowable	adj	48	0.000
taubensee	PERSON	22	0.000
similary	(typo)	3	0.000
recommitment	noun	12	0.000
perorations	noun	3	0.000
orenk	PERSON	2	0.000
malarkey	PERSON	34	0.000
gherardo	PERSON	5	0.000
dcis	ORGANIZATION	3	0.000
...	...	...	...
merritt	PERSON	149	-0.054
echelon	noun	97	-0.058
plugging	verb	265	-0.058
normalcy	noun	170	-0.063
lovell	PERSON	238	-0.066
provisionally	adv	74	-0.068
sails	noun	364	-0.075
rekindled	verb	292	-0.081
sublime	adj	182	-0.090
afflicts	verb	168	-0.116
stan	PERSON	994	-0.132

Table 1: Ranking Result (Single-word)

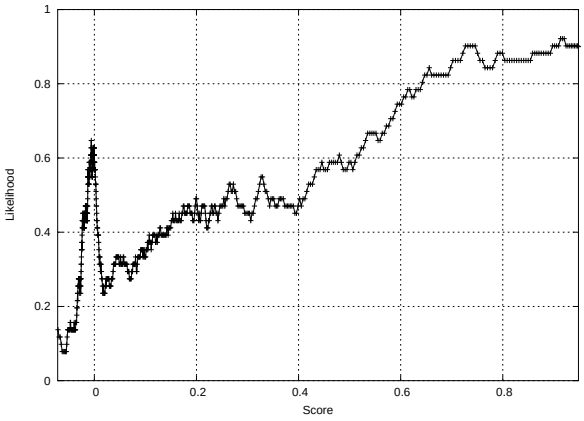


Figure 2: Relationship of the score and the likelihood of being a Named Entity (Single-word). The horizontal axis shows the score of a word. The vertical axis shows the likelihood of being a NE. One can see that the likelihood of NE increases as the score of a word goes up. However there is a huge peak near the score zero.

0.6), we can discover rare Named Entities with 90% accuracy. However, one can notice that there is a huge peak near the score $sim(w) = 0$. This means that many Named Entities still remain in the lower score. Most such Named Entities only appeared in one newspaper. Named Entities given a score less than zero were likely to refer to a completely different entity. For example, the word "Stan" can be used as a person name but was given a negative score, because this was used as a first name of more than 10 different people in several overlapping periods.

Also, we took a look at highly ranked words which are not Named Entities as shown in Table 2. The words "carseats", "tiremaker", or "neurotripic" happened to appear in a small number of articles. Each of these articles and its comparable counterparts report the same event, but both of them use the same word probably because there was no other succinct expression to paraphrase these rare words. This way these three words made a high spike. The word "officeholders" was misrecognized due to the

Word	Type	Freq.	Score
carseats	noun	17	0.9121
tiremaker	noun	21	0.8766
officeholders	noun	101	0.8053
neurotrophic	adj	11	0.7850
mishandle	verb	12	0.7369

Table 2: Errors (Single-word)

	Words	NEs
All words	966	462 (48%)
$sim(w) \geq 0.6$	102	92 (90%)
$sim(w) \leq 0$	511	255 (50%)

Table 3: Obtained NEs (Single-word)

	Word pairs	NEs
All word pairs	810	60 (7%)
$sim(w) \geq 0.05$	27	11 (41%)
$sim(w) \leq 0$	658	30 (5%)

Table 4: Obtained NEs (Multi-word)

repetition of articles. This word appeared a lot of times and some of them made the spike very sharp, but it turned out that the document frequency was undesirably inflated by the identical articles. The word "mishandle" was used in a quote by a person in both articles, which also makes a undesirable spike.

4.2 Multi-word Experiment

In the multi-word experiment, the accuracy of the obtained Named Entities was lower than in the single-word experiment as shown in Table 4, although correlation was still found between the score and the likelihood. This is partly because there were far fewer Named Entities in the test data. Also, many word pairs included in the test data incorrectly capture a noun phrase boundary, which may contain an incomplete Named Entity. We think that this problem can be solved by using a chunk of words instead of two consecutive words. Another notable example in the multi-word ranking is a quoted word pair from the same speech. Since a news article sometime quotes a person's speech literally, such word pairs are likely to appear at the same time in both newspapers. However, since multi-word expressions are much more varied than single-word ones, the overall frequency of multi-word expressions is lower, which makes such coincidence easily stand out. We think that this kind of problem can be alleviated to some degree by eliminating completely identical sentences from comparable articles.

The obtained ranking of word pairs are listed in Table 5. The relationship between the score of word pairs and the likelihood of being Named Entities is plotted in Figure 3.

5 Conclusion and Future Work

In this paper we described a novel way to discover Named Entities by using the time series distribution

Word	Type	Freq.	Score
thai nation	ORG.	82	0.425
united network	ORG.	31	0.290
government open	-	87	0.237
club royale	ORG.	32	0.142
columnist pat	-	81	0.111
muslim minister	-	28	0.079
main antenna	-	22	0.073
great escape	PRODUCT	32	0.059
american black	-	38	0.051
patrick swayze	PERSON	112	0.038
finds unacceptable	-	19	0.034
mayor ron	PERSON	49	0.032
babi yar	LOCATION	34	0.028
bet secret	-	97	0.018
u.s. passport	-	58	0.017
thursday proposed	-	60	0.014
atlantic command	ORG.	30	0.013
prosecutors asked	-	73	0.011
unmistakable message	-	25	0.010
fallen hero	-	12	0.008
american electronics	ORG.	65	0.007
primary goal	-	138	0.007
beach boys	ORG.	119	0.006
amnon rubinstein	PERSON	31	0.005
annual winter	-	43	0.004
television interviewer	-	123	0.003
outside simpson	-	76	0.003
electronics firm	-	39	0.002
sanctions lifted	-	83	0.001
netherlands antilles	LOCATION	29	0.001
make tough	-	60	0.000
permanent exhibit	-	17	0.000

Table 5: Ranking Result (Multi-word)

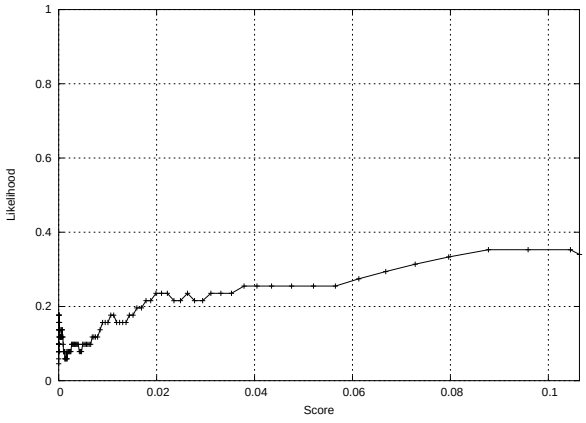


Figure 3: Relationship of the score and the likelihood of being a Named Entity (Multi-word). The horizontal axis shows the score of a word. The vertical axis shows the likelihood of being a NE.

of names. Since Named Entities in comparable documents tend to appear synchronously, one can find a Named Entity by looking for a word whose chronological distribution is similar among several comparable documents. We conducted an experiment with several newspapers because news articles are generally sorted chronologically, and they are abundant in comparable documents. We confirmed that there is some correlation between the similarity of the time series distribution of a word and the likelihood of being a Named Entity.

We think that the number of obtained Named Entities in our experiment was still not enough. So we expect that better performance in actual Named Entity tagging can be achieved by combining this feature with other contextual or lexical knowledge, mainly used in existing Named Entity taggers.

6 Acknowledgments

This research was supported in part by the Defense Advanced Research Projects Agency as part of the Translingual Information Detection, Extraction and Summarization (TIDES) program, under Grant N66001-001-1-8917 from the Space and Naval Warfare Systems Center, San Diego, and by the National Science Foundation under Grant ITS-00325657. This paper does not necessarily reflect the position of the U.S. Government.

References

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL/EACL 2001*.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of EMNLP 1999*.

Satoshi Sekine and Hitoshi Isahara. 2000. IREX: IR and IE evaluation-based project in Japanese. In *Proceedings of LREC 2000*.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of LREC 2002*.

Yusuke Shinyama and Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of International Workshop on Paraphrasing 2003*.

Tomek Strzalkowski and Jin Wang. 1996. A self-learning universal concept spotter. In *Proceedings of COLING 1996*.

Roman Yangarber, Winston Lin, and Ralph Grishman. 2002. Unsupervised learning of generalized names. In *Proceedings of COLING 2002*.