Using Collocation Statistics in Information Extraction

Dekang Lin
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada R3T 2N2
lindek@cs.umanitoba.ca
and
Nalante, Inc.
7 Blackwood Bay, Winnipeg, Manitoba, Canada
lindek@nalante.com

INTRODUCTION

Our main objective in participating MUC-7 is to investigate and experiment with the use of collocation statistics in information extraction. A collocation is a habitual word combination, such as "weather a storm", "file a lawsuit", and "the falling yen". Collocation statistics refers to the frequency counts of the collocational relations extracted from a parsed corpus. For example, out of 6577 instances of "addition" in a corpus, 5190 was used as the object of "in". Out of 3214 instances of "hire", 12 of them take "alien" as the object.

We participated in two tasks: Named Entity and Coreference. In both tasks, the input text is processed in two passes. During the first pass we use the parse trees of input texts, combined with collocation statistics obtained from a large corpus, to automatically acquire or enrich lexical entries which are then used in the second pass.

COLLOCATION DATABASE

We define a collocation to be a dependency triple that consists of three fields:

```
(word, relation, relative)
```

where the `word` field is a word in a sentence, the `relative` field can either be the modifiee or a modifier of `word`, and the `relation` field specifies the type of the relationship between `word` and `relative` as well as their parts of speech.

For example, the dependency triples extracted from the sentence "I have a brown dog" are:

```
(have V:subj:N I)       (I N:r-subj:V have)
(have V:comp1:N dog)    (dog N:r-comp1:V have)
(dog N:jnab:A brown)    (brown A:r-jnab:N dog)
(dog N:det:D a)         (a D:r-det:N dog)
```

The identifiers for the dependency types are explained in Table 1.

Table 1: Dependency types

Label	Relationship between:
`N:det:D`	a noun and its determiner
`N:jnab:A`	a noun and its adjectival modifier
`N:nn:N`	a noun and its nominal modifier
`V:comp1:N`	a verb and its noun object
`V:subj:N`	a verb and its subject
`V:jvab:A`	a verb and its adverbial modifier

We used MINIPAR, a descendent of PRINCIPAR [2], to parse a text corpus that is made up of 55-million-word Wall Street Journal and 45-million-word San Jose Mercury. Two steps were taken to reduce the number of errors in the parsed corpus. Firstly, only sentences with no more than 25 words are fed into the parser. Secondly, only complete parses are included in the parsed corpus. The 100 million word text corpus is parsed in about 72 hours on a Pentium 200 with 80MB memory. There are about 22 million words in the parse trees.

Figure 1 shows an example entry in the resulting collocation database. Each entry contains of all the dependency triples that have the same `word` field. The dependency triples in an entry are sorted first in the order of the part of speech of their `word` fields, then the `relation` field, and then the `relative` field.

The symbols used in Figure (1) are explained as follows. Let `X` be a multiset. The symbol ‖`X`‖ stands for the number of elements in `X` and |`X`| stands for the number of distinct elements in `X`. For example,

a. ‖`(review, V:comp1:N, acquisition)`‖ is the number of times "acquisition" is used as the object of the verb "review".

b. ‖`(review, *, *)`‖ is the number of dependency triples in which the `word` field is "review" (which can be a noun or a verb).

c. ‖`(review, V:jvab:A, *)`‖ is the number of times [$_V$ review] is pre-modified by an adverb.

d. |`(review, V:jvab:A, *)`| is the number of distinct adverbs that were used as a pre-modifier of [$_V$ review].

e. ‖`(*, *, *)`‖ is the total number of dependency triples, which is twice the number of dependency relationships in the parsed corpus.

f. ‖`(review, N)`‖ is the number of times the word "review" is used as a noun.

g. ‖`(*, N)`‖ is the total number of occurrences of nouns.

h. |`(*, N)`| is the total number of distinct nouns that

```
review    8514                                   ||(review, * ,*)||
  V      1424
        V:subj:N        789     179
                administration  5
                appeals court   2
                ... ...                          relation-field
        V:jvab:A        101     39               |(review, V:jvab:A, *)|
                briefly         2                ||(review, V:jvab:A, *)||
                carefully       7
                formally        2
                ... ...
        V:comp1:N       1353    384              relative-field
                account         7
                acquisition     3                ||(review, V:comp1:N, acquisition)||
                action          5
                activity        2
                ... ...                          part of speech
  N      1576                                    ||(review, N)||
        N:r-subj:V      239     118
                affect          8
                become          2
                ... ...
        N:r-comp1:V     525     157
                approve         3
                avoid           5                "review" has been used as the objects
                await           3                of these verbs in the corpus
                ... ...
        N:nn:N          241     85
                admission       2
                bank            2                these nouns were used as a prenominal
                book            4                modifier of "review"
                ... ...
        N:jnp:P         76      9
                by              26
                for             28
                ... ...
        N:jnab:A        518     182
                administrative  5
                annual          12
                antitrust       10
                ... ...
```

The label "word-field" stands above the first line and points to "review".

Figure 1: An example entry in the Collocation Database

i. $\|$`(review, *)`$\|$ is the total number of occurrences of the word "review" (used as any category) in the parsed corpus.

NAMED ENTITY RECOGNITION

Our named entity recognizer is a finite-state pattern matcher, which was developed as part University of Manitoba MUC-6 effort. The pattern matcher has access to both lexical items and surface strings in the input text. In MUC-7, we extended the earlier system in two ways:

- We extracted recognition rules automatically from the collocation database to augment the manually coded pattern rules.
- We treated the collocational context of words in the input texts as features and used a Naive-Bayes classifier to categorized unknown proper names, which are then inserted into the systems lexicon.

A collocational context of a proper name is often a good indicator of its classification. For example, in the 22-million-word corpus, there are 33 instances where a proper noun is used as a prenominal modifier of "managing director". In 26 of the 33 instances, the proper name was classified as an organization. In the remaining 7 instances, the proper name was not classified. Therefore, if an unknown proper name is a prenominal modifier of "managing director", it is likely to refer to an organization. We extracted 3623 such contexts in which the frequency of one type of proper names is much greater (as defined by a rather arbitrary threshold) than the frequencies of other types of proper names. If a proper name occurs in one of these contexts, we can then classify it accordingly. This use of the collocation database is equivalent to automatic generation of classification rules. In fact, some of the collocational contexts are equivalent to pattern-matching rules that were manually coded in the system.

There are only a small number of collocational contexts in which the classification of a proper name can be reliably determined. In most cases, a clear decision cannot be reached based on a single collocational context. For example, among 1504 objects of "convince", 49 of them were classified as organizations, and 457 of them were classified as persons. This suggests that if a proper name is used as the object of "convince", it is likely that the name refers to a person. However, there is also significant probability that the name refers to an organization. Instead of making the decision based on this single piece of evidence, we collect from the input texts all the collocational contexts in which an unknown proper names occurred. We then classify the the proper name with a naive Bayes classifier, using the the set of collocation contexts as features.

The naive Bayes classifier uses a table to store the frequencies of proper name classes in collocational contexts. Sample entries of the frequency table are shown in Table 2. Each row in the table represents a collocation feature. The first column is a collocation feature. Words with this feature have been observed to occur at position X in the second column. The third to fifth columns contain the frequencies of different proper name classes.

Let C be a class of proper name (C is one of LOC, ORG, or PER). Let F_i be a collocation feature. Classification decision is made by find the class C that maximizes $\prod_{i=1}^{k} P(F_i|C)P(C)$,

Table 2: Frequency of Collocation Features

Collocation Feature	Context Pattern	Frequency Counts		
		LOC	ORG	PER
`control|N:r-comp1:V`	to control X	9	87	39
`control|N:r-gen:N`	X's control	14	14	54
`control|N:r-nn:N`	the X control	6	0	0
`control|N:r-subj:V`	X to control	10	99	307
`control|N:subj:N`	X is the control	0	3	0
`convene|N:r-comp1:V`	to convene X	0	5	0
`convene|N:r-subj:V`	X to convene	0	10	18
`convention|N:r-gen:N`	X's convention	0	4	0
`convention|N:r-nn:N`	the X convention	5	23	5

where $F_1, F_2, \ldots F_k$ are the features of an unknown proper name. The probability $P(F_i|C)$ is estimated by m-estimates [5], with $m = 1$ and $p = \frac{1}{|CF|}$ as the parameters, where CF is the set of collocation features:

$$P_m(F_i|C) = \frac{\|F_i, C\| + \frac{1}{|CF|}}{\sum_{f \in CF} \|f, C\| + 1}$$

where $\|F_i, C\|$ denotes the frequency of words that belong to C in the context represented by f.

Example: The walkthrough article contains several occurrences of the word "Xichang" which is not found in our lexicon. The parser extracted the following set of collocation contexts from the formal testing corpus:

1. "the Xichang base", where Xichang is used as the prenominal modifier of "base" (`base|N:nn:N`);
2. "the Xichang site", where Xichang is used as the prenominal modifier of "site" (`site|N:nn:N`);
3. "the site in Xichang", from which two features are extracted:
   - the object of "in" (`in|P:pcomp:N`);
   - indirect modifier of "site" via the preposition "in" (`site|N:pnp-in:N`).

The frequencies of the features are shown in Table 3. These features allowed the naive Bayes classifier to correctly classify "Xichang" as a locale.

Automatically acquiring lexical information on the fly is an double edged sword. On the one hand, it allows classification of proper names that would otherwise be unclassified. On the other hand, since there is no human confirmation, the correctness of the automatically acquired lexical items cannot be guaranteed. When incorrect information is entered into the lexicon, a single error may propagate to many places. For example, during the development of our system, a combination

Table 3: Frequencies of features of "Xichang"

Collocation Feature	Frequency Counts LOC	ORG	PER
base|N:nn:N	77	19	0
site|N:nn:N	26	16	34
in|P:pcomp:N	35641	15630	0
site|N:pnp-in:N	7	0	0

of parser errors and the naive Bayes classification caused the word "I" to be added into the lexicon as a personal name. During the second pass, 143 spurious personal names were generated.

Our NE evaluation results are shown in Table 4. The "pass1" results are obtained by manually coded patterns in conjunction with the classification rules automatically extracted from the collocation database. With the naive Bayes classification, the recall is boosted by 6 percent while the precision is decreased by 2% with an overall increase of F-measure by 2.67.

Table 4: Evaluation results of the named entity task

	Precision	Recall	F-measure
pass1	89%	79%	83.70
official	87%	85%	86.37

COREFERENCE

Our coreference recognition subsystem used the same constraint-based model as our MUC-6 system. This model consists of an integrator and a set of independent modules, such as syntactic patterns (e.g., copula construction and appositive), string matching, binding theory, and centering heuristics. Each module proposes weighted assertions to the integrator. There are two types of assertions. An equality assertion states that two noun phrases have the same referent. An inequality assertion states that two noun phrases must not have the same referent. The modules are allowed to freely contradict one another, or even themselves. The integrator use the weights associated with the assertions to resolve the conflicts. A discourse model is constructed incrementally by the sequence of assertions that are sorted in descending order of their weights. When an assertion is consistent with the current model, the model is modified accordingly. Otherwise, the assertion is ignored and the model remains the same.

One of the important factors to determine whether or not two noun phrases may refer to the same entity is their semantic compatibility. A personal pronoun must refer to a person. For example, the pronoun "it" may refer to an organization, an artifact, but not a person. A "plane" may refer to an aircraft. A "disaster" may refer to a crash. In MUC-6, we used the WordNet to determine the semantic compatibility and similarity between two noun phrases. However, without the ability to determine the intended sense of a word in the input text, we had to say that all senses are possible.[1] The problem with this approach is that the WordNet, like any other general purpose lexical resource, aims at providing broad-coverage. Consequently, it includes many usages of words that are very rare in our domain of interest. For example, one of the 8 potential senses of "company" in WordNet 1.5 is a "visitor/visitant", which is a hyponym of "person". This usage of the word practically never happens in newspaper articles. However, its existence prevents us to make assertions that personal pronouns like "she" cannot co-refer with "company".

In MUC-7, we developed a word sense disambiguation (WSD) module, which removes some of the implausible senses from the list of potential senses. It does not necessarily narrows down the possible senses of a word instance to a single one, however.

Given a polysemous word w in the input text, we take the following steps to narrow down the possibilities for its intended meaning:

1. Retrieve collocational contexts of w from the parse trees of the input text.
2. For each collocational context of w, retrieve its set of collocates, i.e., the set of words that occurred in the same collocational context. Take the union of all the sets of collocates of w.
3. Take the intersection of the union and the set of similar words of w which are extracted automatically with the collocational database [4]. We call the words in the intersection selectors.
4. Score the set of potential senses of w by computing the similarities between senses of w and senses of the selectors in the WordNet [3]. Remove the senses of w that received a score less than 75% of the highest score.

Example: consider the word "fighter" in the following context in the walkthrough article:

> ... in the multibillion-dollar deals for *fighter* jets.

WordNet lists three senses of "fighter":

- combatant, battler, disrupter
- champion, hero, defender, protector
- fighter aircraft, attack aircraft

The disambiguation of this word takes the following steps:

1. The parser recognized that "fighter" was used as the prenominal modifier of "jet".

Table 5: Collocates of "fighter" as prenominal modifier of "jet"

Word	Freq	LogL	Word	Freq	LogL
fighter	80	449.56	NUM	212	160.15
ORG	187	59.56	air force	13	56.28
passenger	17	51.93	Airbus	10	44.18
Lear	6	37.79	Harrier	5	33.62
PROD	14	30.08	-bound	3	22.68
Concorde	4	22.22	Mirage	4	20.02
Avianca	3	15.93	widebody	3	15.66
stealth	4	10.43	turbofan	2	10.35
MiG	2	10.35	KAL	2	9.23
series	5	8.69	cargo	4	8.30
Aeroflot	2	8.16	four-engine	1	7.55
Delta	3	7.53	steering	2	7.09
CANADIENS	2	6.34	water	6	6.23
NUM-passenger	1	6.17	Dragonair	1	6.17
BLACKHAWKS	2	5.98	Skyhawk	1	5.65
Egyptair	1	5.65	transport	3	5.63
trainer	2	5.50	Coast guard	3	5.43
Advanced Tactical Fighter	1	5.31	reconnaissance	2	5.12
Qantas	1	5.05	Pan American	1	5.05
training	3	4.97	United Express	1	4.85
Gulfstream	1	4.85	Swissair	1	4.69
PSA	1	4.69	ANA	1	4.69
ground attack	1	4.54	NUM-seat	1	4.21
Alitalia	1	4.12	Lufthansa	1	3.96
PAL	1	3.89	KLM	1	3.89
NUM Syrian	1	3.76	whirlpool	1	3.03

2. Retrieve words from the collocation database that were also used as the prenominal modifier of "jet" (shown in Table 5). Freq is the frequency of the word in the context, LogL is the log likelihood ratio between the word and the context [1].

3. Retrieve the similar words of "fighter" from an automatically generated thesaurus:

   jet 0.15; guerrilla 0.14; aircraft 0.12; rebel 0.11; bomber 0.11; soldier 0.11; troop 0.10; plane 0.10; missile 0.09; force 0.09; militia 0.09; helicopter 0.09; leader 0.08; civilian 0.07; faction 0.07; pilot 0.07; airplane 0.07; insurgent 0.07; commander 0.06; tank 0.06; airliner 0.05; militant 0.05; marine 0.05; transport 0.05; reconnaissance 0.05; prisoner 0.05; artillery 0.05; army 0.05; stealth 0.05; victim 0.05; terrorist 0.05; weapon 0.04; rocket 0.04; resistance 0.04; rioter 0.04; gunboat 0.04; collaborator 0.04; assailant 0.04; thousand 0.04; gunman 0.04; sympathizer 0.04; radio 0.04; submarine 0.04; attacker 0.04; youth 0.04; camp 0.04; refugee 0.04; dependent 0.04; combat 0.04; mechanic 0.04; demonstrator 0.04; personnel 0.04; movement 0.04; gunner 0.04; territory 0.04

   The number after a word is the similarity between the word and "fighter". The intersection of the similar word list and the above table consists of:

   combat 0.04; reconnaissance 0.05; stealth 0.05; transport 0.05;

4. Find a sense of "fighter" in WordNet that is most similar to senses of "combat", "reconnaissance", "stealth" or "transport". The "fighter aircraft" sense of "fighter" was selected.

We submitted two sets of results in MUC-7:

- the "nowsd" result in which the senses of a word are chosen simply by choosing its first two senses in the WordNet.
- the official result that employs the above word sense disambiguation algorithm.

The results are summarized in Table 6. Although the difference between the use of WSD and the baseline is quite small, it turns out to be statistically significant. In some of the 20 input texts that were scored in coreference evaluation, the WSD module did not make any difference. However, whenever there was a difference it was always an improvement. It is also worth noting that, with WSD, both the recall and precision are increased.

Table 6: Coreference recognition results

	Precision	Recall	F-measure
nowsd	62.7%	57.5%	60.0
official	64.2%	58.2%	61.1

[1] In hindsight, we probably should have just used the first sense listed in the WordNet for each word.

CONCLUSIONS

The use of collocational statistics greatly improved the performance of our named entity recognition system. Although collocation-based Word Sense Disambiguation lead only to a small improvement in coreference recognition, the difference is nonetheless statistically significant.

ACKNOWLEDGEMENTS

This research is supported by NSERC Research Grant OGP121338 and a research contract awarded to Nalante Inc. by Communications Security Establishment.

REFERENCES

[1] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, March 1993.

[2] Dekang Lin. Principle-based parsing without overgeneration. In *Proceedings of ACL–93*, pages 112–120, Columbus, Ohio, 1993.

[3] Dekang Lin. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACL/EACL-97*, pages 64–71, Madrid, Spain, July 1997.

[4] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL '98*, pages 768–774, Montreal, Canada, August 1998.

[5] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.