Structual Ambiguity and Conceptural Information Retrieval

Mathis Huey-chyun Chen and Jason J. S. Chang
Department of Computer Science
National Tsing Hua University
Hsinchu 30043, Taiwan
{ dr804311, jschang }@cs.nthu.edu.tw
telephone: +886-35-731069 fax: +886-35-723694

Abstract

Many researches found lexical preferences to be critical in resolving attachment ambiguity [WFB 90][FBK 82][MP 80]. Most notably, information from Verb-Obj-Prep-Noun structures (*VOPN*) has been used to show that *LA* is very effective in the resolution of PP-attachment ambiguity [HR 93].

We investigated extensions to the lexical association strategy. The extensions include using conceptual association and acquiring the association information from different kind of lexical relations not limited to relations in *VOPN* structures. We refer to this approach as *DeepAttach*. Thus, it is possible to take information from all kinds of syntactical structures as long as they are *alternations* of a common deep structure [PU93] related to that implied by the intended attachment.

A collection of sense-disambiguated sentences serves as the source of conceptual relations. No pre-processing is done to find the conceptual relations in these sentences. Instead, information retrieval technique is used to retrieve conceptually most relevant sentences using the words from the ambiguous structure as query. The prepositional phrase is then attached in favor of the constituent that has more conceptual presence in the ranked retrieved sentences. An experiment was implemented to embody the idea. The result shows that 75% of PP's in 260 *VOPN* structures can be attached correctly, when simple lexical relevance was considered.

1. Introduction

Many researches found lexical preferences to be critical in resolving attachment ambiguity. Jenson and Binot [JB 87] propose to use dictionary definition for disambiguation. [HR 93] describe how co-occurrence of verbs and nouns with prepositions in corpus can be used as an indication of lexical preference. However, dictionary text especially the definitions are typically uneven in their coverage and inconsistent in the usage of words, while lexical co-occurrence suffers from sparseness of data. Resnik and Hearst (1994) find that using class information did not yield improved performance over lexical association due to multiplicity of classes for a word and lack of disambiguation.

Our proposal is to use sense-disambiguated, sizable body of text as the source of lexical preference. Thus, for example, in the sentence 'He woke up to find the house on fire,' one of the word senses of the noun 'fire' occurs frequently in the context of the words which are conceptually similar to 'house.' This is evidence of a conceptual association of the direct object 'house' with the prepositional object 'fire.' Unlike in previous proposals [HR 93][RH 93], these co-occurrences of concepts need not in the Verb-Object-Preposition relation.

2. The DeepAttach Algorithm

For simplicity, the *verb*, *direct object*, *preposition*, and *head of prepositional object* of each testing sentence are fed to the prepositional phrase disambiguator. Category of Longman LEXICON [LO 92] is used as our semantic tag set. Given the sentence 'He woke up to find the house on fire,' the word 'fire' has four different semantic tags, and the other two words have only one tag, respectively.

He	woke	up	to	find	the	house	on	fire.
				N1361		Da007		Hc075, Hd120, Hh245, Jh212

Thus we look into our database which is composed of 25,000 automatically sense-disambiguated sentences [KC 95] to retrieve and rank sentences that are semantically similar to the testing sentence. Only the pairs of verb/preposition, direct-object/preposition, verb/direct-object, verb/prepositional-object, and direct-object/prepositional-object of the testing sentence are examined. Note that we only utilize the part-of-speech and semantic information of each word of the 25,000 sentences in the database. Syntactical relations such as verb/object relations are not used.

We use the above example to illustrate our idea. Conceptually, we use the following query to retrieve relavant sentences from the database. Notice that this query pattern contains both the tags of detailed level such as 'Bg082' and of coarse level such as 'Bg'. Therefore, the query will not be too narrow; some sentences will always be retrieved.

```
[('find' and 'on') or ('house' and 'on') ] or
[(N1361 and 'on') or (Da007 and 'on')] or
[(N1361 and (Hc075 or Hd120 or Hh245 or Jh212)) or
(Da007 and (Hc075 or Hd120 or Hh245 or Jh212))] or
[(N1 and (Hc or Hd or Hh or Jh)) or
(Da and (Hc or Hd or Hh or Jh))]
```

To process the query more efficiently, we create indexes of the database similar to the inverted file used in information retrieval.

word$_a$	word$_b$	sense$_a$	sense$_b$	tag$_a$	tag$_b$	position$_a$	position$_b$	sentence#
I	catch	Gh280	De098	pron	v	1	2	1
I	a	Gh280	Nd098	pron	det	1	3	1
catch	a	De098	Nd098	v	det	2	3	1
I	fish	Gh280	Ea017	pron	n	1	4	1
catch	fish	De098	Ea017	v	n	2	4	1
a	fish	Nd098	Ea017	det	n	3	4	1
I	yesterday	Gh280	Lh225	pron	n	1	5	1
catch	yesterday	De098	Lh225	v	n	2	5	1
a	yesterday	Nd098	Lh225	det	n	3	5	1
fish	yesterday	Ea017	Lh225	n	n	4	5	1

The query returns a list of sentences. We show a few examples in the following.

```
1. exact match at the detailed level:
The house is on fire. (subject/predicative-pp)
```

```
    Da007         Hc075
2. match at both levels:
The house is burning! (subject/verb)
    Da007    Hc076
3. exact match at the coarse level:
The building burnt down to ashes. (subject/verb and subject/p-object)
    Da003    Hc076          Hc084
```

It leads to a quick conclusion that the pair 'house-fire' is better than other possible pairs such as 'woke-fire,' and 'find-fire.' And the prepositional phrase 'on fire' is more likely to attach to the preceding noun phrase rather than to the verb. Besides, there is one by-product that the word 'fire' can be disambiguated as to have the word sense of 'Hc075' in the testing sentence.

To be quantitative, a ranking function must be designed to denote the degree of *resemblance and compatibility* both in structure and in semantics.

3. Experiment and Results

An experiment was implemented to embody the idea. Currently, only simple lexical relevance was considered.

To perform an outside test, we select 260 sentences not in the database and manually mark the verb, direct object, preposition, and head of prepositional object for easy processing. Two human judges are arranged to examine the results. It shows that 75% of PP's in these 260 sentences can be correctly attached.

4. Analysis and Discussion

Examining those errors, we found some interesting issues.

Our original idea is to blend the data extracted from different structures to solve the problem of data deficiency. However, it will also introduce some kind of noises. For example, our model incorrectly attaches the PP to the Direct-Object for 'acquire knowledge of language by studying' because *knowledge, language, and studying* are highly-related in their senses.

We observe that the resulting attachment implies an intrinsic Subject/Object or Subject/Predicate relation but most retrieved evidences from the database are of incompatible Predicate/Object structure. This seems to suggest that the compatibility between structures of the query and the database entry should be considered.

Another issue that we should consider is the diversity of the prepositions. Different prepositions have different tendencies of attachment. therefore present different degree of difficulty. See Figure 1. for details. This suggests that each preposition should be handled differently.

5. Ongoing Work

In this preliminary experiment, the ranking function of a sentence is simply based on counting. There are many possible ways to refine it. Furthermore, we only take lexical and semantic information into account. The polarity and the distance between the lexical items are left out. We are currently constructing a more elaborated model that will handles these two aspects and the structural compatibility as well.

Besides, we assume that the sense-disambiguated database is 100% correct. In fact, the precision is only about 80%. This inaccuracy in the sense tags obviously has a negative effect in the results of our experiments. We are simultaneously improving the accuracy of sense disambiguation.

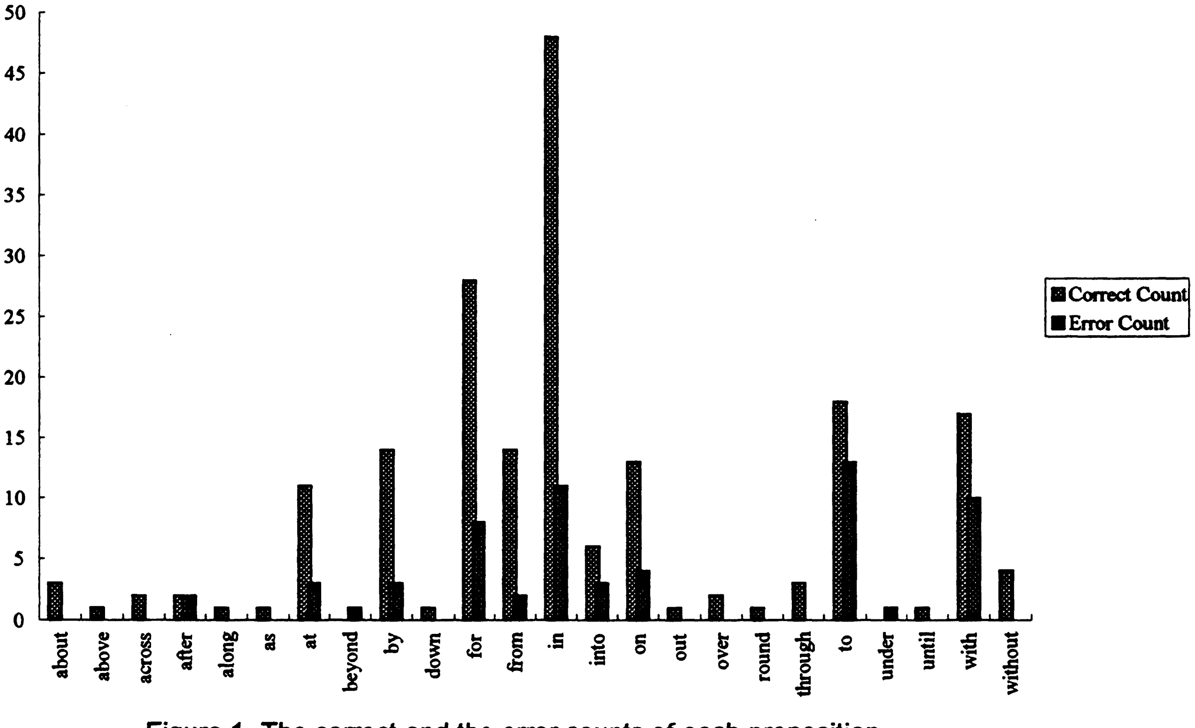

Figure 1. The correct and the error counts of each preposition.

References

[BR 94] E. Brill and P. Resnik, A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation. COLING-94.

[CB 95] M. Collins and J. Brooks, Prepositional Phrase Attachment through a Backed-Off Model. ACL 3rd Workshop on Very Large Corpora, 1995.

[FB 92] W. B. Frakes and R. Baeza-Yates, Information Retrieval - Data Structures & Algorithms, Prentice-Hall, 1992.

[FBK 82] M. Ford, J. Bresnan, and R. Kaplan, A Competence Based Theory of Syntactic Closure, in The Mental Representation of Grammatical Relations, MIT Press, 1982.

[HR 93] D. Hindle and M. Rooth, Structural Ambiguity and Lexical Relations. Computational Linguistics, 19(1):103-120, 1993.

[HI 87] G. Hirst, Semantic Interpretation and the Resolution of Ambiguity, Cambridge University Press, 1987. ISBN 0 521 32203 0.

[JB 87] K. Jenson and J. Binot, Disambiguating Prepositional Phrase Attachments by Using On-Line Dictionary Definitions, Computational Linguistics, 13(3-4): 251-260, 1987.

[KC 95] S. J. Ker and Jason J.S. Chang. Simultaneous Resolution of Word Alignment and Sense Ambiguity - An approach based on thesaurus classes, submitted to PACLIC-10, 1995.

[LO 92] Longman, Longman English-Chinese Dictionary of Contemporary English, Published by Longman Group (Far East) Ltd., Hong Kong, 1992.

[MC 92] Tom McArthur, Longman Lexicon of Contemporary English, Published by Longman Group (Far East) Ltd., Hong Kong, 1992.

[MA 80] Marcus and P. Mitchel, A Theory of Syntactic Recognition for Natural Language. MIT Press , 1980.

[ME 93] J. J. Mei, et al. Tongyici Cilin (A word-tree of symnonyms), Tong Hua Publishing, Taipei, 1993 (traditional Chinese edition of a simplified Chinese edition published in 1984).

[PU 93] J. Pustejovsky, Sabine Bergler, and Peter Anick, Lexical Semantic Techniques for Corpus Analysis. Computational Linguistics, 19(2):331-358, 1993.

[RA 94] A. Ratnaparkhi, J. Reynar, and S. Roukos, A Maximum Entropy Model for Prepositional Phrase Attachment. ARPA Workshop on Human Language Technology, 1994.

[RH 93] P. Resnik and M. A. Hearst. Structural Ambiguity and Conceptual Relations. In Proceedings of ACL Workshop on Very Large Corpora, 1993.

[UR 95] N. Uramoto, Automatic Learning of Knowledge for Example-Based Disambiguation of Attachment, TMI-95.

[WFB 90] G. Whittemore, K. Ferrara, and H. Brunner. Empirical Study of Predicative Powers of Simple Attachment Schemes, ACL-90.