SUBJECT-DEPENDENT CO-OCCURRENCE AND WORD SENSE DISAMBIGUATION

Joe A. Guthrie, Louise Guthrie, Yorick Wilks, and Homa Aidinejad*

Computing Research Laboratory
Box 30001
New Mexico State University
Las Cruces, NM 88003-0001

ABSTRACT

We describe a method for obtaining subject-dependent word sets relative to some (subject) domain. Using the subject classifications given in the machine-readable version of Longman's Dictionary of Contemporary English, we established subject-dependent co-occurrence links between words of the defining vocabulary to construct these "neighborhoods". Here, we describe the application of these neighborhoods to information retrieval, and present a method of word sense disambiguation based on these co-occurrences, an extension of previous work.

INTRODUCTION

Word associations have been studied for some time in the fields of psycholinguistics (by testing human subjects on words), linguistics (where meaning is often based on how words co-occur with each other), and more recently, by researchers in natural language processing (Church and Hanks, 1990; Hindle and Rooth, 1990; Dagan, 1990; McDonald et al., 1990; Wilks et al., 1990) using statistical measures to identify sets of associated words for use in various natural language processing tasks.

One of the tasks where the statistical data on associated words has been used with some success is lexical disambiguation. However, associated word sets gathered from a general corpus may contain words that are associated with many different senses. For example, vocabulary associated with the word "bank" includes "money", "rob", "river" and "sand". In this paper, we describe a method for obtaining subject-dependent associated word sets, or "neighborhoods" of a given word, relative to a particular (subject) domain. Using the subject classifications of Longman's Dictionary of Contemporary English (LDOCE), we have established subject-dependent co-occurrence links between words of the defining vocabulary to construct these neighborhoods. We will describe the application of these neighborhoods to information retrieval, and present a method of word sense disambiguation based on these co-occurrences, an extension of previous work.

CO-OCCURRENCE NEIGHBORHOODS

Words which occur frequently with a given word may be thought of as forming a "neighborhood" of that word. If we can determine which words (i.e. spelling forms) co-occur frequently with each word sense, we can use these neighborhoods to disambiguate the word in a given text.

Assume that we know of only two of the classic senses of the word *bank*:

1) A repository for money, and
2) A pile of earth on the edge of a river.

We can expect the "money" sense of *bank* to co-occur frequently with such words

* Present address: Mathematics Department, University of Texas at El Paso, El Paso, Tx 79968

as "money", "loan", and "robber", while the "river" sense would be more frequently associated with "river", "bridge", and "earth". In order to disambiguate "bank" in a text, we would produce neighborhoods for each sense, and intersect them with the text, our assumption being that the neighborhood which shared more words with the text would determine the correct sense. Variations of this idea appear in (Lesk, 1986; McDonald, et al., 1990; Wilks, 1987; 1990; Veronis and Ide, 1990).

Previously, McDonald and Plate (McDonald et al., 1990; Schvaneveldt, 1990) used the LDOCE definitions as their text, in order to generate co-occurrence data for the 2,187 words in the LDOCE control (defining) vocabulary. They used various methods to apply this data to the problem of disambiguating control vocabulary words as they appear in the LDOCE example sentences. In every case however, the neighborhood of a given word was a co-occurrence neighborhood for its spelling form over all the definitions in the dictionary. Distinct neighborhoods corresponding to distinct senses had to be obtained by using the words in the sense definition as a core for the neighborhood, and expanding it by combining it with additional words from the co-occurrence neighborhoods of the core words.

SUBJECT-DEPENDENT NEIGHBORHOODS

The study of word co-occurrence in a text is based on the cliche that "one (a word) is known by the company one keeps". We hold that it also makes a difference *where* that company is kept: since a word may occur with different sets of words in different contexts, we construct word neighborhoods which depend on the subject of the text in question. We call these, naturally enough, "subject-dependent neighborhoods".

A unique feature of the electronic version of LDOCE is that many of the word sense definitions are marked with a subject field code which tells us which subject area the sense pertains to. For example, the "money"-related senses of *bank* are marked *EC* (Economics), and for each such main subject heading, we consider the subset of LDOCE definitions that consists of those sense definitions which share that subject code. These definitions are then collected into one file, and co-occurrence data for their defining vocabulary is generated. Word x is said to co-occur with word y if x and y appear in the same sense definition; the total number of times they co-occur is denoted as f_{xy}.

We then construct a 2,187 x 2,187 matrix in which each row and column corresponds to one word of the defining vocabulary, and the entry in the xth row and yth column represents the number of times the xth word co-occurred with the yth word. (This is a symmetric matrix, and therefore it is only necessary to maintain half of it.) We denote by f_x the total number of times word x appeared. While many statistics may be used to measure the relatedness of words x and y, we used the function

$$r(x,y) = \frac{f_{xy}}{f_x + f_y - f_{xy}}$$

in this study. We choose a co-occurrence neighborhood of a word x from a set of closely related words. We may choose the ten words with the highest relatedness statistic, for instance.

Neighborhoods of the word "metal" in the category "Economics" and "Business" are presented below:

Table 1. Economics neighborhood of *metal*

Subject Code EC = Economics				
metal	idea	coin	them	silver
	real	should	pocket	gold
	well	him		

Table 2. Business neighborhood of *metal*

Subject Code BU = Business				
metal	bear	apparatus	mouth	inside
	spring	entrance	plate	brass
	tight	sheet		

In this example, the neighborhoods reflect a fundamental difference between the two subject areas. Economics is a more theoretical subject, and therefore its neighborhood contains words like "idea", "gold", "silver", and "real", while in the more practical domain of Business, we find the words "brass", "apparatus", "spring", and "plate".

We can expect the contrast between subject neighborhoods to be especially great for words with senses that fall into different subject areas. Consider the actual neighborhoods of our original example, *bank*.

Table 3. Economics neighborhood of *bank*

Subject Code EC = Economics				
bank	account	cheque	money	by
	into	have	keep	order
	out	pay	at	put
	from	draw	an	busy
	more	supply	it	safe

Table 4. Engineering neighborhood of *bank*

Subject Code EG = Engineering				
bank	river	wall	flood	thick
	earth	prevent	opposite	chair
	hurry	paste	spread	overflow
	walk	help	we	throw
	clay	then	wide	level

Notice that even though we included the twenty most closely related words in each neighborhood, they are still unrelated or disjoint, although many of the words which appear in the lists are indeed suggestive of the sense or senses which fall under that subject category. In LDOCE, three of the eleven senses of *bank* are marked with the code *EC* for Economics, and these represent the "money" senses of the word. It is a quirk of the classification in LDOCE that the "river" senses of bank are not marked with a subject code.

This lack of a subject code for a word sense in LDOCE is not uncommon, however, and as was the case with *bank*, some word senses may have subject codes, while others do not. We label this lack of a subject code the "null code", and form a neighborhood of this type of sense by using all sense definitions without code as text. This "null code neighborhood" can reveal the common, or "generic" sense of the word.

The twenty most frequently occurring words with *bank* in definitions with the null subject code form the following neighborhood:

Table 5. Null Code neighborhood of *bank*

Subject Code NULL = no code assigned				
bank	rob	river	account	lend
	overflow	flood	money	criminal
	lake	flow	snow	cliff
	police	shore	heap	thief
	borrow	along	steep	earth

It is obvious that approximately half of these words are associated with our two main senses of *bank*-but a new element has crept in: the appearance of four out of eight words which refer to the *money* sense ("rob", "criminal", "police", and "thief") reveal a sense of *bank* which did not appear in the EC neighborhood. In the null code definitions, there are quite a few references to the potential for a bank to be robbed.

Finally, for comparison, consider a neighborhood for *bank* which uses *all* the LDOCE definitions (see McDonald et al., 1990; Schvaneveldt, 1990; Wilks et al., 1990):

Table 6. Unrestricted neighborhood of *bank*

Subject Code All				
bank	account	bank	busy	cheque
	criminal	earn	flood	flow
	interest	lake	lend	money
	overflow	pay	river	rob
	safes	and	thief	wall

Only four of these words ("bank", "earn", "sand", and "thief") are not found in

the other three neighborhoods, and the number of words in the intersection of this neighborhood with the Economics, Engineering, and Null neighborhoods are: six, four, and eleven, respectively. Recalling that the Economics and Engineering neighborhoods are disjoint, this data supports our hypothesis that the subject-dependent neighborhoods help us to distinguish senses more easily than neighborhoods which are extracted from the whole dictionary.

There are over a hundred main subject field codes in LDOCE, and over three-hundred sub-divisions within these. For example, "medicine-and-biology" is a main subject field (coded "MD"), and has twenty-two sub-divisions such as "anatomy" and "biochemistry". These main codes and their sub-divisions constitute the only two levels in the LDOCE subject code hierarchy, and main codes such as "golf" and "sports" are not related to each other. Currently, we use only the main codes when we are constructing a subject-dependent neighborhood. But even this division of the definition text is fine enough so that, given a word and a subject code, the word may not appear in the definitions which have that subject code at all.

To overcome this problem, we have adopted a restructured hierarchy of the subject codes, as developed by Slator (1988). This tree structure has a node at the top, representing all the definitions. At the next level are six fundamental categories such as "science" and "transportation", as well as the null code. These clusters are further subdivided so that some main codes become sub-divisions of others ("golf" becomes a sub-division of "sports", etc.). The maximum depth of this tree is five levels.

If the word for which we want to produce a neighborhood appears too infrequently in definitions with a given code, we travel up the hierarchy and expand the text under consideration until we have reached a point where the word appears frequently enough to allow the neighborhood to be constructed. The worst case scenario would be one in which we had traveled all the way to the top of the hierarchy and used all the definitions as the text, only to wind up with the same co-occurrence neighborhoods as did McDonald and Plate (Schvaneveldt, 1990; Wilks et al., 1990)!

There are certain drawbacks in using LDOCE to construct the subject-dependent neighborhoods, however: the amount of text in LDOCE about any one subject area is rather limited, is comprised of a control vocabulary for dictionary definitions only, and uses sample sentences which were concocted with non-native English speakers in mind.

In the next phase of our research, large corpora consisting of actual documents from a given subject area will be used, in order to obtain neighborhoods which more accurately reflect the sorts of texts which will be used in applications. In the future, these neighborhoods may replace those constructed from LDOCE, while leaving the subject code hierarchy and various applications intact.

WORD SENSE DISAMBIGUATION

In this section, we describe an application of subject-dependent co-occurrence neighborhoods to the problem of word sense disambiguation. The subject-dependent co-occurrence neighborhoods are used as building blocks for the neighborhoods used in disambiguation. For each of the subject codes (including the null code) which appear with a word sense to be disambiguated, we intersect the corresponding subject-dependent co-occurrence neighborhood with the text being considered (the size of text can vary from a sentence to a paragraph). The intersection must contain a pre-selected minimum number of words to be considered. But if none of the neighborhoods intersect at greater than this threshold level, we replace the neighborhood N by the neighborhood $N(1)$, which consists of N together with the first word from each neighborhood of words in N, using the same subject code. If necessary, we add the second most strongly associated word for each of the words in the original neighborhood N, forming the neighbor-

hood *N(2)*. We continue this process until a subject-dependent co-occurrence neighborhood has intersection above the threshold level. Then, the sense or senses with this subject code is selected. If more than one sense has the selected code, we use their definitions as cores to build distinguishing neighborhoods for them. These are again intersected with the text to determine the correct sense.

The following two examples illustrate this method. Note that some of the neighborhoods differ from those given earlier since the text used to construct these neighborhoods includes any example sentences which may occur in the sense definitions. Those neighborhoods presented earlier ignored the example sentences. In each example, we attempt to disambiguate the word "bank" in a sentence which appears as an example sentence in the Collins COBUILD English Language Dictionary. The disambiguation consists of choosing the correct sense of "bank" from among the thirteen senses given in LDOCE. These senses are summarized below.

bank(1) : [ ] : land along the side of a river, lake, etc.

bank(2) : [ ] : earth which is heaped up in a field or garden.

bank(3) : [ ] : a mass of snow, clouds, mud, etc.

bank(4) : [AU] : a slope made at bends in a road or race-track.

bank(5) : [ ] : a sandbank in a river, etc.

bank(6) : [AU] : to move a car or aircraft with one side higher than the other.

bank(7) : [ ] : a row, especially of oars in an ancient boat or keys on a typewriter.

bank(8) : [EC] : a place in which money is kept and paid out on demand.

bank(9) : [MD] : a place where something is held ready for use, such as blood.

bank(10) : [GB] : (a person who keeps) a supply of money or pieces for payment in a gambling game.

bank(11) : [ ] : break the bank is to win all the money in bank(10).

bank(12) : [EC] : to put or keep (money) in a bank.

bank(13) : [EC] : to keep ones money in a bank.

Example 1. The sentence is "The aircraft turned, banking slightly."
The neighborhoods of "bank" for the five relevant subject codes are given below.

Table 7. Automotive neighborhood of *bank*

Subject Code AU = Automotive				
bank	make	go	up	move
	so	they	high	also
	round	car	side	turn
	road	aircraft	slope	bend
	safe			

Table 8. Economics neighborhood of *bank*

Subject Code EC = Economics				
bank	have	it	person	out
	into	take	money	put
	write	keep	pay	order
	another	paper	draw	supply
	account	safe	sum	cheque

Table 9. Gambling neighborhood of *bank*

Subject Code GB = Gambling				
bank	person	use	money	piece
	play	keep	pay	game
	various	supply	chance	

Table 10. Medical neighborhood of *bank*

Subject Code MD = Medicine and Biology				
bank	something	use	place	hold
	medicine	ready	blood	human
	origin	organ	store	hospital
	treatment	product	comb	

Table 11. Null Code neighborhood of *bank*

Subject Code NULL = No code assigned				
bank	game	earth	stone	boat
	river	bar	snow	lake
	sand	shore	mud	framework
	flood	cliff	heap	harbor
	ocean	parallel	overflow	clerk

The AU neighborhood contains two words, "aircraft" and "turn", which also appear in the sentence. Note that we consider all forms of turn (turned, turning, etc.) to match "turn". Since none of the other neighborhoods have any words in common with the sentence, and since our threshold value for this short sentence is 2, AU is selected as the subject code. We must now decide between the two senses which have this code.

At this point we remove the function words from the sense definitions and replace each remaining word by its root form. We obtain the following neighborhoods.

Table 12. Words in sense 4 of *bank*

Definition bank(4)				
slope	make	bend	road	so
they	safe	car	go	round

Table 13. Words in sense 6 of *bank*

Definition bank(6)			
car	aircraft	move	side
high	make	turn	

Since bank(4) has no words in common with the sentence, and bank(6) has two ("turn" and "aircraft"), bank(6) is selected. This is indeed the sense of "bank" used in the sentence.

Example 2. The sentence is "We got a bank loan to buy a car." The original neighborhoods of "bank" are, of course, the same as in Example 1. The threshold is again 2. None of the neighborhoods has more than one word in common with the sentence, so the iterative process of enlarging the neighborhoods is used. The AU neighborhood is expanded to include "engine" since it is the first word in the AU neighborhood of "make". The first word in the AU neighborhood of "up" is "increase", so "increase" is added to the neighborhood. If the word to be added already appears in the neighborhood of "bank", no word is added.

On the fifteenth iteration, the EC neighborhood contains "get" and "buy". None of the other neighborhoods have more than one word in common with the sentence, so EC is selected as the subject code. Definitions 8, 12, and 13 of bank all have the EC subject code, so their definitions are used as cores to build neighborhoods to allow us to choose one of them. After twenty-three iterations, bank(8) is selected.

Experiments are underway to test this method and variations of it on large numbers of sentences so that its effectiveness may be compared with other disambiguation techniques. Results of these experiments will be reported elsewhere.

FURTHER APPLICATIONS

Several applications of subject-dependent neighborhoods in addition to word-sense disambiguation are being pursued, as well. For information retrieval, previously constructed neighborhoods relevant to the subject area can be used to expand a query and the target (titles, key words, etc.) to include more words in the intersection, and improve both recall and precision. Another application is the determination of the subject area of a text. Since the effectiveness of searching for key words to determine the topic of a text is limited by the choice of the particular list of key words, and the fact that the text may use synonyms or refer to the concept the key word represents without using it (for example by using a pronoun in its place), we could look for word associations (thereby involving more words in the process and making it less vulnerable to the above problems),

rather than simply searching for key words indicative of a topic. Neighborhoods of words in the text could be constructed for each of the six fundamental categories, and intersected with the surrounding words in the text. After choosing the category with the greatest intersection, we would then traverse the subject code tree downward to arrive at a more specific code, stopping at any point where there is not enough data to allow us to choose one code over the others at that level. Once a subject code is selected for a text, it could be used as a context for word-sense disambiguation.

CONCLUSION

Although the words in the LDOCE definitions constitute a small text (almost one million words, compared with the mega-texts used in other co-occurrence studies), the unique feature of subject codes which can be used to distinguish many definitions, and LDOCE's small control vocabulary (2,187 words) make it a useful corpus for obtaining co-occurrence data. The development of techniques for information retrieval and word-sense disambiguation based on these subject-dependent co-occurrence neighborhoods is very promising indeed.

ACKNOWLEDGEMENTS

This research was supported by the New Mexico State University Computing Research Laboratory through NSF Grant No. IRI-8811108. Grateful acknowledgement is accorded to all the members of the CRL Natural Language Group for their comments and suggestions.

REFERENCES

Church, Kenneth W., and Patrick Hanks (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16, 1, pp.22-29.

Dagan, Ido, and Alon Itai (1990). Processing Large Corpora for Reference Resolution. *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, Helsinki, Finland, 3, pp.330-332.

Hindle, Donald, and Mats Rooth (1990). Structural Ambiguity and Lexical Relations. *Proceedings of the DARPA Speech and Natural Language Workshop.*

Lesk, Michael E. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *Proceedings of the ACM SIGDOC Conference*, Toronto, Ontario.

McDonald, James E., Tony Plate, and Roger W. Schvaneveldt (1990). Using Pathfinder to extract semantic information from text. In R. W. Schvaneveldt (ed.), *Pathfinder Associative Networks: Studies in Knowledge Organization.* Norwood, NJ: Ablex.

Schvaneveldt, Roger W. (1990). *Pathfinder Associative Networks: Studies in Knowledge Organization.* New Jersey: Ablex.

Slator, Brian M. (1988). Constructing Contextually Organized Lexical Semantic Knowledge-bases. *Proceedings of the Third Annual Rocky Mountain Conference on Artificial Intelligence (RMCAI-88)*, Denver, CO, pp.142-148.

Veronis, Jean., Nancy Ide (1990). Very Large Neural Networks for Word-sense Disambiguation. COLING '90, 389-394.

Wilks, Yorick A., Dan C. Fass, Chengming Guo, James E. McDonald, Tony Plate, and Brian M. Slator (1987). A Tractable Machine Dictionary as a Resource for Computational Semantics. Memorandum in Computer and Cognitive Science, MCCS-87-105, Computing Research Laboratory, New Mexico State University. In Branimir Boguraev and Ted Briscoe (eds.), *Computational Lexicography for Natural Language Processing.* Harlow, Essex, England: Longman Group Limited.

Wilks, Yorick A., Dan C. Fass, Chengming Guo, James E. McDonald, Tony Plate, and Brian M. Slator (1990). Providing Machine Tractable Dictionary Tools. Journal of Machine Translation, 2. Also to appear in Theoretical and Computational Issues in Lexical Semantics , J. Pustejovsky (Ed.)