

WORD SENSE DISAMBIGUATION OF MODALS (ext. abstract)

1. introduction

The work reported on in this abstract concerns the question of the ambiguity of modal verbs in English. Most modals are ambiguous in that they have more than one distinct sense. The questions addressed in this paper are:

- What are the senses of modals?
- What are the disambiguation factors? Are they the same as for lexical disambiguation (as reported in among others Ide and Véronis 1999 and Stevenson and Wilks 2001)?
- What is the context needed for a successful disambiguation?

In this abstract, I will limit myself to the modal *must*. Its senses can be clearly defined and the disambiguation factors are relatively clear. In the final version of the paper, other modals such as *may*, *can*, and *should* will be discussed as well.

2. modals and meaning

In this abstract I will assume that modals have two distinct senses, exemplified in (...) below.

- (1) a. John must go to school.
b. John must be in school.

When queried, most speakers of English agree that the modal *must* has a different meaning in (1a) than it has in (1b). The meaning of *must* in (1a) is commonly referred to as an instance of **deontic** modality (the term **root** modality is also used, see e.g., Coates 1983), while (1b) uses the **epistemic** sense of *must*. While finer distinctions are possible (for instance, Palmer 1990 distinguishes between deontic and **dynamic** modality), the present study will assume that English modals are two-way ambiguous (have two distinct senses that need to be disambiguated). For the moment, this disregards those modals that have more meanings (such as *can* which also has an **ability** reading).

The question is: if a modal such as *must* is ambiguous, why do people assign without hesitation a deontic reading to (1a) and an epistemic reading to (1b)? It is quite possible to construct scenarios where sentence (1a) receives an epistemic interpretation and (1b) a deontic one. These readings are not salient, however, and the question is: why not? What about the context makes *must* in (1a) a deontic modal, but an epistemic one in (1b)?

3. analysis of the corpus

Before proceeding to the WSD stage, the corpus was analyzed by hand according to a number of features. The results are reported in De Haan 2003. Two such features will be exemplified here, verbal construction and person of subject. Other factors which may possibly be relevant include negation, subjective vs. objective reading of the modal, and semantic status of the main verb. It is not obvious that knowing the precise meaning of the main verb would help, but it is conceivable that a Vendlerian division of accomplishment, achievement, stative, and activity verbs might be helpful. Such a task has not been added here, but it could easily be done.

There were 520 sentences containing the modal *must* in the Switchboard corpus, of which 11 were ungrammatical or irrelevant to the present study. These were discarded. Of the remaining 509 sentences, 66 were deontic, 412 epistemic, and 31 were indeterminate, which means that even within context it is not possible to assign a definitive interpretation to the modal.

The 509 sentences were analyzed on their verbal collocations. An example is shown in (2):

- (2) That's the only place I was sore, and I thought, well, I must not be doing them right ... [S272]

This sentence has the verbal form *must be V-ing* (where V signifies the main verb) and is coded accordingly. The full list of distributions is shown in Table 1.

Table 1
Occurrences of *must* with verbal complements in the Switchboard corpus

	Deontic	Epistemic	Indeterminate
<i>Be +V- part.</i>	4		1
<i>V</i>	54	55	7
<i>Have V</i>	3	34	4
<i>Have been V</i>		70	
<i>Be</i>	2	195	5
<i>Be + V-ing</i>		9	1
<i>Have got + V-ed</i>		2	
<i>No V</i>	3	5	13
<i>Have + V-ed</i>		41	
<i>Have been +V-ing</i>		1	

As can be seen from the data in Table 1, this parameter serves very well to disambiguate between the two senses of *must*. There are only two constructions that are heavily ambiguous, namely *must V* (with an almost even split between the two senses) and *must* without accompanying main verb. Perhaps unsurprisingly, the latter construction had the highest number of indeterminate cases.

The second parameter looked at is the person of the subject. The results are shown in Table 2 below. Note that no distinction was made between second person singular and plural because the form *you* is ambiguous in itself. As is to be expected most persons occur more with epistemic modals than with deontic ones due to the overwhelming number of epistemic sentences in the corpus. The one exception to this is the first person singular, which has an overwhelming preference for deontic *must*. Impersonal subjects include constructions like *it must be* or *there must be*.

Table 2
Correlation of person and modality in the Switchboard corpus

	Deontic	Epistemic	Indeterminate
<i>1 SG</i>	39	14	1
<i>2 SG/PL</i>	10	67	5
<i>3 SG</i>	6	166	13
<i>1 PL</i>	2	13	1
<i>3 PL</i>	7	59	10
<i>No overt subject</i>	1	29	0
<i>Impersonal subject</i>	1	64	1
Total	66	412	31

These numbers show that these two parameters are possible disambiguation criteria and a system was designed based on these features.

4. the WSD results

The corpus used for this study is the Switchboard corpus, a corpus of spoken American English, which was tagged with POS tags according to the CLAWS C7 scheme. Under this scheme, all modals, excluding *catenative* modals such as *ought to*, receive the POS tag **VM**. This is irrespective of its modal sense, and is just meant as a syntactic tag.

The first 100 occurrences of *must* in the Switchboard that had an unambiguous meaning were hand tagged with their respective sense (this excludes the indeterminate cases). Thus an epistemic instance was coded with the tag **VME** and a deontic one was coded with **VMD**. This was used as a training corpus and the system then proceeded to analyze the verbal structure of each sentence based on its POS analysis. The system was therefore not provided with the numbers of Table 1 and 2, but was expected to deduce the information from the linguistic material itself. This turned out to be impossible with the subject, since the

system did not have access to the syntactic structure of the sentence at this point. The subject was hand coded at the second stage.

After all errors were smoothed out (this includes programming errors and POS errors) the system, based on just the first criterion, verbal collocation, already had a successful disambiguation rate of about 81%. With the second feature, subject person, added in, the success rate climbed marginally to about 83% (this parameter was only used on the *must V* and *must* without main verb constructions, since these were the constructions with the highest level of ambiguity). The main problem lies as expected in the *must V* construction, which proves very difficult to disambiguate automatically. Also, the system assigns an interpretation to any sentence which is indeterminate. Since these interpretations cannot be checked for accuracy, they bring down the success rate, albeit marginally.

The system is currently rerunning the examples, but now Treebank information is added to give more syntactic information. It is still too soon to report on the results of these experiments (but they will be included in the final paper). Current work is also focused on finding the combination of criteria that will yield the highest degree of accuracy.

5. preliminary conclusions

The results from this experiment show that a disambiguating system based on a single parameter, verbal collocation, already drastically improves on randomly assigning a meaning. For an automated system that has to make a determination very quickly this might suffice, but it will still make an error 20% of the time on average. For a system which aims to mimic human behavior, more is obviously needed.

An additional problem is that *must* behaves differently in different forms of the language. According to the data in Biber *et al.* 1999, *must* is predominately epistemic in the spoken language, but predominately deontic in the written language. This is confirmed in my data: the verb *must* is epistemic in 79% of all cases in the Switchboard corpus, but only 16% of the time in the written Brown corpus. *Must* is the only modal for which this is true. It is not quite clear how to deal with that. It is of course possible to construct a training corpus based on a mixture of sentences from the spoken and written language and use these numbers as a basis. This disregards the fact that the choice of speech style itself is a disambiguating factor and can be used as such. In this case we need two sets of data, one for each style and the situation determines which set will be used.

References

- Biber, Douglas, *et al.* (eds.). 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Coates, Jennifer. 1983. *The semantics of the modal auxiliaries*. London: Croom Helm.
- De Haan, Ferdinand. 2003. *must*-constructions. Manuscript, University of Arizona.
- Fellbaum, Christiane (ed.). 1999. *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press.
- Ide, Nancy; Jean Véronis. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics* 24(1), 1-40.
- Palmer, Frank R. 1990. *Modality and the English modals, second edition*. London: Longman.
- Stevenson, Mark; Yorick Wilks. 2001. The Interaction of Knowledge Sources in Word Sense Disambiguation. *Computational Linguistics* 27(3), 321-49.
- SWITCHBOARD CORPUS. *Linguistic Data Consortium*.