Reading Distinction in MT*

Pius ten Hacken
Eurotra
Onderzoeksinstituut voor Taal en Spraak
University of Utrecht
Trans 10, 3512 JK Utrecht
email: tenhacken@hutruu59.bitnet

Abstract

In any system for Natural Language Processing having a dictionary, the question arises as to which entries are included in it. In this paper, I address the subquestion as to whether a lexical unit having two senses should be considered ambiguous or vague with respect to them. The inadequacy of some common strategies to answer this question in Machine Translation (MT) systems is shown. From a semantic conjecture, tests are developed that are argued to give more consistent and theoretically well-founded results.

1 Introduction

In any system for Natural Language Processing having a dictionary, the question arises which entries are included in it. In this paper, I will assume the environment of a multilingual MT system based on a linguistic analysis and transfer architecture, from which I will derive some argumentation.

The question which entries are included in the dictionary should be answered in two parts. First there is a mapping from graphic words to lexical units (lu's), then a mapping from lu's to readings, each of which is represented in an entry. The former mapping represents a certain level of analysis of the graphic word. It abstracts away from inflection and spelling variation, and, depending on the system's analysis component, may do so as well for productive derivation and compounding, and multi-word units. In this paper I will concentrate on the latter mapping, reading distinction, in a way that does not appeal to a particular choice on the relation between lu and graphic word.

A consistent approach to reading distinction is necessary, because inconsistencies in reading distinction in an MT system will complicate transfer components between a pair of languages, and jeopardize extensibility of the system. A correct solution will save time in development and improve performance. The central question in this area can be formulated as in (1).

(1) Given an lu X and two of its senses S_1 and S_2, is X ambiguous or vague with respect to S_1 and S_2 ?

In (1) a sense of an lu is the meaning the lu has in a certain set of contexts. If the lu is vague, both senses are covered by the same reading. If it is ambiguous, S_1 and S_2 are examples of different readings of the lu, each reading being represented in a single entry.

2 Some common methods

Since every reading distinction creates lexical ambiguity that has to be solved, it seems attractive to use the features expressing the relevant information as criterion for answering (1): An lu is ambiguous between S_1 and S_2 iff there is a feature describing the difference.

If we only take morphological and syntactic features, many intuitively clear cases of ambiguity (e.g. *bank* as financial institution vs. as

*I would like to thank my colleagues at the university and in Eurotra, especially Louis des Tombe and Henk Verkuyl, for their helpful comments.

river verge) cannot be expressed. This will lead to problems in transfer or in generation. On the other hand, these features will cause unwanted distinctions as well, e.g. French *fonctionnaire* (civil servant) with gender masculine or feminine, and *kneel* with past tense *kneeled* or *knelt*. It makes no sense to try to disambiguate these.

The use of semantic features to define ambiguity should be rejected for similar reasons. First, we have to determine a fixed set of features a priori, since otherwise no answers to questions of reading distinction evolve. This imposes an artificial upper bound on reading distinction. Moreover, the availability of a certain feature does not mean that it has to be assigned in all cases. We will certainly have a feature expressing the male/female contrast, but it is not desirable to create two readings of *parent* accordingly, leading to a translation of *parents* into something meaning *mother(s) and/or father(s)*.

Alternatively, we could argue that, since translation is the goal, it should be the criterium for reading distinction as well, answering (1): An lu having two senses is ambiguous iff there are different translations for the two senses. Leaving aside the non-trivial problem of determining whether there are different translations, we have to admit that there are cases of exceptional distinctions in one language, e.g. *fleuve* vs. *rivière* in French, meaning river ending in a sea or in another river respectively. These distinctions will influence all other dictionaries in the system, in the sense that e.g. English *river* and Dutch *rivier* become ambiguous, and there are two translation rules between them. If we restrict our attention to a limited group of languages, e.g. the languages in the system, the system becomes difficult to extend, since adding a new language from outside this group will affect all existing dictionaries. Otherwise there is a conceptual problem, since it will never be possible to decide that an lu is vague, unless we know all languages of the world. Instead, cases of exceptional distinctions, *bilingual ambiguities*, are best handled in transfer between the two languages, because they really are translation problems.

Summarizing, taking the means (features) or the goal (translation) as a criterium for reading distinction results in decisions that cause various practical problems and are intuitively incorrect. Furthermore these strategies detach the notion of *reading* from *meaning*, which is theoretically undesirable.

Taking only intuitions as our guide will link reading to meaning, but if even trained intuitions of lexicographers do not prevent inconsistencies, as can be seen in many published dictionaries, there is not much hope of reaching consistency, unless we manage to find some support for the intuitions.

3 A semantic method

The tests I will propose here to decide on reading distinction are based on monolingual meaning, and yield a substantially greater degree of consistency than direct, unaided intuitions. It is based on the following conjecture.

(2) There is a set of processes P, that, given a single occurrence of an lu, can stretch the actual meaning of the lu in the context to the boundaries of the reading the lu has, but not beyond.

In order to be able to check the results, we will first consider some tests where specific processes in P are used, as applied to some intuitively clear cases of ambiguity (e.g. *bank* as financial institution vs. river verge) and vagueness (e.g. *elephant* as Indian elephant vs. African elephant). Then the scope of these tests will be expanded to other cases.

A well-known test evolving from (2) is based on conjunction. Lakoff (1970) proposed a test where anaphoric *so* in the second clause of a conjunction refers back to an antecedent containing the lu for which the question of reading distinction arises, as in (3).

(3) a. John went to a bank this morning, and so did Mary.
b. John saw an elephant, and so did Mary.

The question to be asked in this case is whether the sentence is semantically normal when the anaphor is interpreted in the other sense than its antecedent. Clearly, (3a) is strange in this interpretation, whereas (3b) is normal, confirming that *bank* but not *elephant* is ambiguous in the relevant way (cf. Cruse (1986) on the use of semantic normality judgements). Other anaphors, e.g. *one*, *there* can be used as well. The answers are more reliable in case of an antecedent containing less lexical material outside the lu in question. In (3a), the antecedent of *so* is *go to a bank*, and ambiguity might be claimed to arise from the verb. Using *one* instead of *so* takes away this possibility, which is especially relevant in less clear cases.

Other processes use quantifiers. One, based on Wiggins (1971), uses universal quantification. It is exemplified in (4).

(4) a. All banks in this town are safe.
b. All elephants in this zoo are old.

The question to be asked here is whether *all X* can be interpreted as *all* S_1 or *all* S_2, or only as *all* S_1 *and all* S_2. Whereas (4a) can mean either that there is no danger of flooding or that bank-robbers are effectively discouraged, and it is odd when used to mean both, (4b) can only be used to predicate over both African and Indian elephants in the zoo that they are old. A variant using negation in the same way is discussed by Kempson & Cormack (1981).

A slightly different test can be performed with a universal quantifier somewhat remote from the relevant lu, as in (5).

(5) a. Every town has a bank.
b. Every zoo has an elephant.

The question to be asked here is whether the X (bank/elephant) has to be interpreted in the same sense for every Y (town/zoo). In a similar way numerals can be used as in (6), and coordination as in (7), requiring the same question.

(6) a. This town has two banks.
b. This zoo has two elephants.

(7) a. John and Mary went to a bank this morning.
b. John and Mary saw an elephant this morning.

Summarizing, there are three main classes of processes in P behaving as in (2). The first one refers to two elements from the extension of the lu, one of them by an anaphor, as in (3). The second one refers to the full extension of a reading at once, as in (4). The third one refers to a group of elements in the extension, exploiting distributivity, as in (5)–(7). Each class is associated with a different question the answer of which determines whether an analysis as ambiguity or as vagueness is correct. There are various realizations of test sentences for each class, some of which are subject to independently motivated constraints. In a natural way an intuitively appealing definition of *reading* evolves as in (8).

(8) A reading of an lu is a coherent group of senses, the boundaries of which cannot be crossed by a single occurrence of the lu without losing semantic normality.

4 The tests in actual use

In the previous section, semantic tests were shown to give correct answers in cases where we can check them. This proves that we should not immediately reject the tests. The reason we need them however, is that there are many cases where unaided intuition is not sufficiently determinate, so that conflicts on the correct analysis might arise.

A well-known problem area is the analysis of privative oppositions, where one of the senses is more general and includes the other one. Both *dog* and *lion* have senses *animal belonging to a particular species of mammals* and *male specimen of that species*. According to Kempson (1980) they are both vague with respect to these senses, but Zwicky & Sadock (1975) claim that *dog* but not *lion* is ambiguous. Applying various tests to them we get the following sentences.

(9) a. John has a dog, and Mary has one too.
b. The zoo has a lion, and the circus has one too.

(10) a. All dogs of this breed are short-sighted.
b. All lions in this wild reserve have been killed by poachers.

(11) a. This family has two dogs.
b. This zoo has two lions.

The sentences (9) and (11) cannot lead to a conclusion for independent reasons. Since an individual or a group in the more specific sense of the lu is also an individual or a group in the general sense, the general sense is always available to cover up the opposition. This is not the case when the full extensions are compared, however. Therefore from (10) we can indeed conclude that *dog* is ambiguous and *lion* is not. Both (10a) and (10b) have the general interpretation, but only (10a) also has the more specific one (cf. *..., but not the bitches* vs. *..., *but not the lionesses*).

Another problem that comes up is the construction of test sentences for other syntactic categories than nouns. Although the various processes are most easily demonstrated with nouns, nothing in the theory refers to nouns directly. VP-anaphors, e.g. *so*, can also be used for verbs.

(12) a. John has been running all day, and so has his washing machine.
b. John has been running all day, and so has his dog.

(13) John followed Mary, and Bill did so too.

The sentences in (12) show the ambiguity of *run* between the senses with a human and a machine subject, and the vagueness between senses with a biped and a quadruped subject. For transitive verbs, such as *follow*, having the sense *understand* and *go after*, the result of the test is more disputable, since (13) shows the ambiguity of *follow Mary*, and one could argue that it is due to ambiguity of *Mary*, e.g. between the senses *thinking person* and *spatial object*. Therefore, the use of a non-lexical anaphor, indicated by # in the examples, is to be preferred.

(14) John followed Mary, and Bill # Kate.

It is rather difficult to construct a sentence with a quantifier over the verb comparable to (4) for nouns. Rather, a sentence such as (15) below displays the same distributivity effect as (5), that can also be achieved by coordination as in (16).

(15) All boys followed Mary.
(16) John and Bill followed Mary.

The test sentences for verbs can also be used for adjectives, if they are used predicatively. An example is (17), where *black* is shown to have different readings when used with a concrete object and with *humour*.

(17) Her dress is black, and so is her humour.

For gradable adjectives, a comparison is a basis for constructing a test sentence. Although (17) can be used humoristically, (18) below, illustrating the ambiguity of *fair*, can hardly be interpreted.

(18) Her hair is as fair as the salary she pays her employees.

In general it seems that for gradable adjectives comparison provokes stronger judgements than anaphoric reference by *so*. In some cases, however, one of the senses cannot be used predicatively, and neither of the two processes can be used. An empty anaphor sometimes provides a solution, as in (19), where *economic* is shown to be ambiguous between the senses *relating to the economy* and *not wasteful*.

(19) For many years, he produced economic theories and # cars.

In some languages, there is a lexical anaphor that requires an adjective as its antecedent, e.g. *dito* in Dutch, as illustrated in (20).

(20) Bij hun gouden bruiloft kregen
ze een dito horloge.
(Litt.) 'At their golden wedding got
they a # watch'

Among the remaining problems is the comparison of two senses with big syntactic differences. All test sentences have to be syntactically correct, and syntax does not allow e.g. coordination of a noun and a verb in corresponding positions. In such cases, the semantic part of testing the senses is never arrived at.

5 Conclusion

In this paper, I developed tests to answer the question whether an lu with two senses is to be analyzed as ambiguous or vague with respect to them from the semantic conjecture (2). The tests allow for theoretically well-founded and consistent decisions in many cases. In MT, they determine a proper balance on the cline between what can easily be disambiguated monolingually, and what is useful as a distinction in translation. As such they define the target for monolingual disambiguation, and the class of bilingual ambiguities, that should be treated in transfer. Since the MT environment has only been used in the argumentation, not in the solution proposed, theoretical well-foundedness and consistency evolving from the tests presented here are equally valid in other environments where a monolingual dictionary is used.

References

Cruse, D.A. (1986). *Lexical Semantics*, Cambridge University Press.

Kempson, Ruth (1980). ' Ambiguity and Word Meaning', in: Greenbaum, Sidney, Geoffrey Leech & Jan Svartvik, *Studies in English Linguistics*, Longman, London / New York, p. 7-16.

Kempson, Ruth & Annabel Cormack (1981). 'Ambiguity and Quantification', *Linguistics and Philosophy 4*, p. 259-309.

Lakoff, George (1970). 'A Note on Vagueness and Ambiguity', *Linguistic Inquiry 1*, p. 357-359.

Wiggins, David (1971). 'On sentence-sense, word-sense and difference of word-sense. Towards a philosophical theory of dictionaries.' In: Steinberg, Danny & Leon Jakobovits (ed.). *Semantics*, Cambridge University Press, p. 14-34.

Zwicky, Arnold & Jerrold Sadock (1975). 'Ambiguity tests and how to fail them', in: Kimball, John (ed.). *Syntax and semantics 4*, Academic Press, p. 1-36.