Chapter 11
Senses and Texts

Introduction

Empirical, corpus-based, computational linguistics reaches by now into almost every crevice of the subject, and perhaps pragmatics will soon succumb. Semantics, if we may assume the sense-tagging task is semantic, taken broadly, has shown striking progress and, in Yarowsky’s work (1995) produced very high levels of success in the 90%±, well above the key bench-mark figure of 62% correct sense assignment, achieved at an informal experiment in New Mexico about 1990, in which each word was assigned its first sense listed in LDOCE.

A crucial question in this chapter will be whether recent work in sense-tagging has in fact given us the breakthrough in scale that is now obvious with, say, part-of-speech tagging. Our conclusion will be that it has not, and that the experiments so far, however high their success rates, are not yet of a scale different from those of the previous generation of linguistic, symbolic-AI or connectionist approaches to the very same problem.

A historian of our field might glance back at this point to, say, Small et al. (1988) which covered the AI-symbolic and connectionist traditions of sense-tagging at just the moment before corpus-driven empirical methods began to revive. All the key issues still unsettled are discussed there and the collection showed no naiveté there about the problem of sense resolution with respect only to existing lexicons of senses. It was realised that that task was only meaningful against an assumption of some method for capturing new (new to the chosen lexicon, that is) senses and, most importantly, that although existing lexicons differed, they did not differ arbitrarily much. The book also demonstrated that there was also strong psychological backing for the reality of word senses and for empirical methods of locating them from corpora without any prior assumptions about their number or distribution (e.g. Plate’s work in Wilks et al. 1990, and see also Jorgensen, 1990).

Our purpose in this chapter will be to argue that Kilgarriff’s (1993) negative claims are simply wrong, when he argued that there really are no word senses and hence they cannot be discriminated by computer Yarowsky is largely right in arguing against this, as we shall show, although we have some queries about the details and the interpretation of his claims. Both authors however agree that this is a traditional and important task: one often cited as being, because of the inability of systems
of the past to carry it out, a foundational lacuna in, say, the history of machine translation (MT). It was assumed by many, in that distant period, that if only word-sense ambiguity could be tamed, by the process we are calling sense-tagging, then MT of high quality would be relatively straightforward. Like may linguistic tasks, it became an end in itself, like syntactic parsing, and, now that it is, we would claim, firmly in sight (despite Kilgarriff) it is far less clear that its solution will automatically solve a range of traditional problems like MT. But clearly it would be a generally good tool to have and local triumph if this long-resistant bastion of NLP were to yield.

The Very Possibility of Sense-Tagging

Kilgarriff’s paper (1993) is important because it has been widely cited as showing that the senses of a word, as distinguished in a dictionary such as LDOCE, do not cover the senses actually carried by most occurrences of the word as they appear in a corpus. If his paper does show that, it is very significant indeed, because that would imply that sense-tagging word occurrences in a corpus by means of any lexical data based on, or related to, a machine-readable dictionary or thesaurus is misguided. I want to show that here the paper does not demonstrate any such thing. Moreover, it proceeds by means of a straw-man it may be worth bringing back to life!

That straw-man, Kilgarriff’s starting point, is the ‘bank model’ (BM) of lexical ambiguity resolution, which is established by assertion rather than quotation, though it is attributed to Small, Hirst, and Cottrell as well as the present author. In the BM, words have discrete meanings, and the human reader (like the ideal computer program) knows instantly and effortlessly which meaning of the word applies (Ibid. p. 367), “given that a word occurrence always refers to one or the other, but not both” of the pair of main meanings that a word like ‘bank’ is reputed to have. The main thrust of Kilgarriff’s paper is to distinguish a number of relationships between LDOCE senses that are not discrete in that way, and then to go on to an experiment with a corpus.

But first we should breathe a little life back into the BM straw-man: those named above can look after themselves, but here is a passage from Wilks (1972, p. 12) “.it is very difficult to assign word occurrences to sense classes in any manner that is both general and determinate. In the sentences “I have a stake in this country” and “My stake on the last race was a pound” is “stake” being used in the same sense or not? If “stake” can be interpreted to mean something as vague as “Stake as any kind of investment in any enterprise” then the answer is yes. So, if a semantic dictionary contained only two senses for “stake”: that vague sense together with “Stake as a post”, then one would expect to assign the vague sense for both the sentences above. But if, on the other hand, the dictionary distinguished “Stake as an investment” and “Stake as an initial payment in a game or race” then the answer would be expected to be different. So, then, word sense disambiguation is relative to the dictionary of sense choices available and can have no absolute quality about it”. QED.