Chapter 5: Dimensions of context

49-53

5.1 Introduction

The context of a word is everything that occurs before, during and after a word is uttered including the word itself. Since we cannot take everything into consideration, it is useful to regard previous and future events as lying along a continuum and to consider the role and importance of contextual information as a function of distance from the target word. Some explicit attempts at defining the importance of the context as a function of the distance have been made (Yarowsky and Florian, 2002), but the most common approach is to define the importance in terms of linguistic units surrounding the word, e.g., phrase, clause, sentence, discourse, topic, domain.

The context has at least three different dimensions whose impact on word sense discovery and disambiguation are seldom investigated independently: context size, modality of context[48], and depth of context processing. The easiest one to investigate is the context size. For word senses the size and shape of the context has been found to influence different word categories differently, e.g., nouns often benefit more from a longer context than verbs which often depend on a local context[49].

5.2 Context Size

When looking at the size of the context, it is useful to divide it into three separate size categories:

zero context – the word itself

local context – phrase and clause

global context – sentence, discourse, topic and domain

The size categories are assumed to be independent of the modalities and the depth of processing in each of the modalities. The zero context is sometimes disregarded, but it contains important information about the internal structure of the word, e.g., capitalization, sounds or morphs, which often relate the word to other words with similar meaning. The local context is sometimes defined as a narrow window of 3-5 words centered on the word itself. A narrow window is an approximation of the linguistic concepts, such as phrase and clause, which may not be available without linguistic software. The global context is defined as a window of 25-1000 words centered on the word itself, which fairly well approximates contexts starting from the immediately surrounding sentence to the whole document or domain. It is sometimes useful to define

syntactic context separately as the context provided by the heads and immediate constituents, because the syntactic context may carry information crossing the borders of all three of the context sizes. The syntactic relations which are most important for characterizing the word sense of a word can often be found within the same clause boundaries[49] as the word itself, so in practice they can be regarded as part of the local context.

We wish to point out that the division of context into different sizes is purely a tool for linguistically motivating the experiments, where we extracted different sizes of contexts in order to show their relative influence on word sense disambiguation. One could argue that there are no distinct boundaries between local and global context[50]. There are only more or less influential context features, whose general tendency is that their influence diminishes with increasing distance from the word itself.

5.3 Context Modality

Human language technology systems have typically focused on the "factual" aspect of content analysis. Other aspects, including pragmatics, point of view, and style, have received much less attention. However, to achieve an adequate understanding of a text, these aspects cannot be ignored. (Qu et al., 2004) The two primary modalities for perceiving language are hearing (audition) and seeing (vision). In addition to these two, we learn the meaning of words in context by taste (gustation)1, smell (olfaction) and several types of physical feeling (tactition, thermoception, nociception, equilibrioception and propioception[51]). The main categories make up the five Aristotelian senses. Some animals have at least three more: electroception, magnetoception and echolocation, which to humans may seem as instances of the "sixth" sense. Two of these additional senses, i.e., electroception and magnetoception are used by psychologists in the form of EEG (Electro Encephalo Gram) and MRI (Magnetic Resonance Imaging) to study brain activity. As humans we grow up learning to interpret written information in relation to our senses. If the words in a machine-readable dictionary could be given appropriate and relevant initial values for all of these senses, it would perhaps be possible for a computer to make better generalizations of basic semantic features such as animate and inanimate.

In the future, robots may be conditioned on this kind of life experience. In the meantime, we may have to encode some of the basic lexical semantic features separately for each word in order to bolster computer inferences based on natural language text input. The concept hierarchies of machine-readable thesauruses and ontologies are potential sources for such world-knowledge. Even if WordNet (Fellbaum, 1998)[52] does not give explicit lexical domain relations like "sports" for the words "racket" and "court", the information can be extracted with some additional processing. Similarly, other more fundamental lexical semantic features could be precomputed by inference. Another similar source on a larger scale is the Internet, especially if the effort to create a Semantic Web is successful .

5.4 Depth of Preprocessing

As a separate dimension of context we have the depth of preprocessing. There are two radically different ideological points of view for context preprocessing. Some aim at encoding underlying principles of language and communication and expect language models to emerge as a by-product of applying such principles to a corpus. If this is the goal, it is useful to get by with as little preprocessing as possible relying mainly on machine learning and statistical language models. Some aim at drawing on existing linguistic knowledge and use as much preprocessing as is efficiently available and necessary to solve new problems. We will now outline the existing levels of computational linguistic preprocessing. Looking at the surface form of a word, we find that the original capitalization of a word form is an annotation entered by the author of a document. If the word forms are normalized so that capital letters are turned into lower case, at least one prominent distinction is lost between, e.g., Church as an institution and church as a building. Traditionally, a base form is the form found in a dictionary. Some word forms may have several base forms depending on context. In English the base form is often unique. The ambiguity is mainly between parts-of-speech with the same base form. One notable exception is the analysis of participles, e.g., "a drunk driver/has drunk a lot" with base forms drunk/drink or "was heading south/the newspaper heading is" with base forms head/heading etc. The correct base form can be determined in context as a side-effect of part-of-speech tagging. An intermediate level before full dependency parsing is head syntax, which only indicates in which direction the head word is and what

part of speech the head word is. The main advantage of head syntax is that it avoids attachment ambiguities, e.g., in “the man on the hill with the telescope” the preposition with is tagged as a dependent of some noun to the left, e.g., with N . Full dependency syntax builds parse trees with one head word for each word. Each head word may have several dependents. For a rule-based approach to dependency syntax, see Tapanainen and J¨arvinen (1997), and for a statistical approach, see Samuelsson (2000)[53].

An alternative route would be feature structure-based context descriptions, e.g., using unification, which were popular in the beginning of the 1990s (Carlson and Lind´en, 1987; Gazdar and Mellish, 1989) and which was adopted in the LFG (Kaplan and Bresnan, 1982) and HPSG (Pollard and Sag, 1994) frameworks. However, the wide-coverage FDG parsers (Connexor, 2002b) are now also available with attribute-value feature structure representation of the output. The attribute-value grammars may at the time have seemed too static for capturing usage preferences, such as connotations, and extended usages, such as metaphors. There should at least have been a mechanism in them for learning usage preferences based on observations, but it was not until Abney (1997) that the attribute value graphs received a proper stochastic framework with an algorithm to estimate the parameters. The algorithm was rather heavy to compute. Recently interesting attempts at creating morphological (Creutz and Lagus, 2004)[54] and syntactic (Klein, 2005) parsers using unsupervised learning have been made. The idea is to incorporate general language independent principles in a natural language morphology or syntax discovery algorithm and then try to find a grammar and a lexicon that embody the training material as succinctly as possible. It is likely that the quality of the output will improve as the encoded discovery principles become more specific, even if the goal is to find as general principles as possible. We are aiming at semantic similarity defined as substitutability in context, so we need to study how far we can get with a piecewise representation of the context and the linguistic structures in the context.3 As a summary, computational linguistics currently offers the following high quality domain-independent preprocessing: _ token analysis, i.e., word forms are separated from punctuation marks and fixed expressions are identified. In some languages this phase may also include identifying potential word boundaries. morphological analysis, i.e., dictionary look-up of inflectional tags for word forms. Part-

of-speech tagging is morpho-syntactic analysis disambiguating inflectional tags. syntactic analysis, i.e., immediate constituents and a syntactic head-word are identified for each word in a clause. The nature of the identification may vary from a simple indication of the direction of the head-word to a precise co-indexing, which can serve to build a tree structure or a dependency graph.