CONCEPT-ORIENTED PARSING OF DEFINITIONS

WILLY MARTIN
Free University of Amsterdam
The Netherlands

0. Introduction

Computational lexicology/lexicography currently favours issues related to the acquisition, the representation and the application of lexical knowledge functioning within a NLP-environment. Within the first domain, that of acquisition, one finds e.g. such topics as the extraction of information from corpora and machine readable dictionaries (MRD's). This paper - which is the result of a project in which, next to the author, also M.Reedijk, E.ten Pas and L.Willekens participated -, falls within this first domain as it will explicitly deal with the extraction of (lexical) knowledge from (dictionary) definitions. However, it will be evident that acquisition without a representational framework does not make (much) sense. Furthermore we will also indicate how to use the knowledge obtained.

1. Definitions and Meaning Types

Our starting-point is the fact that words, with regard to their meaning, can be classified into meaning types. Words can have meanings that are predominantly conceptual, collocational, grammatical, figurative, connotative, stylistic and contextual/discursive. A word such as the geological term magma typically has a conceptual meaning only, another one such as bloody (as in 'you bloody fool') typically combines collocational meaning (intensification) with stylistic meaning aspects (very informal), whereas the same word, bloody, e.g. in a sentence like 'I got my bloody foot caught in the bloody chair' (example taken from LDOCE) mainly gets a discursive, a contextual, meaning (functioning as an (emotional) stopword). Different kinds of lexical meaning types require different descriptive treatments. So e.g. terms, showing 'par excellence' conceptual meaning, will require first and foremost conceptual meaning descriptions i.e. concept-oriented definitions. In what follows then we will concentrate on terms and their meaning as expressed in definitions, the typical locus for conceptual meaning information. Accordingly the parser we will present will be concept-oriented.

2. Concept-oriented parsing of terms

The parser under discussion is set up to analyze definitions of medical terms in English. As such it is but one of the components of a system which at the moment consists of a preprocessor, a segmentor, a lexicon, a set of conceptual relations and a parser proper. In order to better understand the approach under discussion we will

- first give a general overview of the overall algorithm
- thereafter globally comment upon those aspects which are most relevant from a computational linguistic point-of-view (as it is impossible, given the amount of

space and time, to give a full and detailed picture of the whole project).

2.1 Overall algorithm

The basic algorithm can be roughly characterized as consisting of the following steps:

a. read definition
b. segment definition
c. look for head of definition
d. check clues
e. look for subhead(s) of definition
f. fill frame subhead(s) taking into account (checks on)
   - coordination
   - clues
   - postmodification
g. fill frame head
h. write sense frame

A typical input reads like this:
"rheumatoid arthritis: a chronic disease of the musculo-skeletal system, characterized by inflammation and swelling of the joints, muscle weakness, and fatigue" (taken from Collins Dictionary of the English Language 1986[2])

The corresponding output looks like

rheumatoid_arthritis:

```
[disease    g_affects     musc_skel_syst]
[disease    has_qual      chronic]
[disease    has_symptom   fatigue]
[disease    has_symptom   weakness]
[disease    has_symptom   inflammation]
[disease    has_symptom   swelling]
  [weakness -   g_affects   muscle]
  [swelling     g_affects   joints]
  [inflammation g_affects   joints]
```

In what follows we will try to make clear the main features (a system leading to) such a result implies.

2.2 Basic features

2.2.1 Input specifications

Up till now we have only dealt with definitions for diseases (terms for nosology concepts). These definitions can be taken from all kinds of sources, e.g. from termbanks or from (terminological) dictionaries. The example given above should make clear that we work with analytical definitions exhibiting all kinds of difficulties in both lexis and syntax (such as structural ambiguities cf. 'inflammation' vs. 'inflammation of the joints').

2.2.2 Lemmatizer-tagger as Front-end

It goes without saying that a lemmatizer-tagger is a basic requirement for the efficient operation of the parser. This way text words (= word forms occurring in the definitional text) can be linked up with the items occurring in the lexicon (see below). For that purpose we use an adapted version of Dilemma (see Martin e.a. 1988 and Paulussen-Martin 1992).

2.2.3 Minimal syntax

After having been lemmatized and tagged, the definition gets split up into smaller parts (segments) by the segmentor. This module is a minimal syntactic processor which, on the basis of categorial information (such as Boolean values for NP compatibility and NP delimitation), delimits word groups in the input string. Unlike other approaches (such as Alshawi 1989) which make use of syntactic pattern matching techniques, syntax is kept to a strict minimum as one of our claims is that

much of what is done (by others) syntactically, can be left out when one disposes of more powerful, i.c. conceptual, knowledge. As a result our input definition now looks as follows (| indicating delimiters, || indicating boundaries):

a chronic disease| of the musculo-skeletal system|, (characterized) | by inflammation| and swelling | of the joint(s) |, muscle weakness |, and fatigue ||.

2.2.4 Conceptual knowledge and calculation

The knowledge banks which form the core of the system are the lexicon and the set of conceptual relations.

A lexical entry, e.g. aids, is a three-place predicate consisting of the actual lexeme, its concept type and its word category. So:

(aids, concept (nosology-concept, aids, [u, u, u, u, u, u]), n).

As one will observe, the second argument, the concept type, consists of a sixtuple, i.c. six unspecificied slots. The parsing of definitions is precisely aimed at filling or specifying these slots.

It is the set of conceptual relations that a concept type may have that determines this specification. At the moment such a relational template for diseases (nosology concepts), somewhat simplified, looks as follows:

nos-concept
g_affects (nos, {macro, micro, funct, embryo})
o_affects (nos, organism)
caused_by (nos, etiology)
has_symptom (nos, finding)
transmitted_by (nos, trans)
has_qual (nos, qual)

In our approach the universe of discourse is split up into 22 interrelated concepttypes, which, as a rule, form homogeneous subsets. At the center of it one finds nosology concepts which show relations with other concepttypes which in their turn may show relations with other concepttypes, which in their turn are related to other concepttypes, etc.

At this point it is important to see that implicit concepts such as nosology concepts, (and so the conceptual meaning of the lexeme aids e.g.) can a.o. be defined/specified by concepts taken from the domain of macro- and micro-anatomy and that, in the given case, the relation between both arguments will be established. In this respect it is crucial for the parser to find the head concept of the definitional phrase. It does so by setting up a syntax-based hypothesis (taking the rightmost noun occurring in front of the first delimiter) and checking it with conceptual knowledge. In case of a definition of aids as

"a group of diseases secondary to a defect in cell-mediated immunity associated with a single newly discovered virus" (taken from Eurodicautom)

in a first instance group will be taken up as head. Afterwards it will be rejected on conceptual grounds, a.o. because of the fact that group is not considered a medical concept. In other cases head shifting will take place because of the fact that the head candidate can not be conceptually specified by its subheads (conceptual incompatibility between the assumed head and its subhead(s)). In the same vein, when being confronted with "classes of phenomena that present great difficulties for all syntactic formalisms (...) [One of], the most important of these being conjunction (...)" (Winograd 1983, 257-258), the parser again will solve (or try to solve) these cases by making use of conceptual information. That, in the case of rheumatoid arthritis, it does not yield parses such as 'swelling of muscle

(weakness)' and that it manages to combine 'joints' both with 'swelling' and 'inflammation' proves it to be fairly successful in this respect. Other examples of conceptual calculation imply the establishment of new concept types out of old ones. 'Throat' e.g., being a macro-anatomical concept, becomes a finding concept when in combination with a qual concept such as in 'sore', this way 'sore throat' can 'fill' a symptom relation with a nosology concept. This example also shows that the lexicon is conceived as an atomic one: concepts are thought of as atoms from which more complex structures can be derived; if the latter are compositional and can be computed however, they are not taken up as such. Other examples of conceptual parsing include the application of rules for PP-attachment. Compare: "a disease characterized by a sense of constriction in the chest" vs. "a discase characterized by a sense of constriction in children". In the former case the PP 'in the chest' will be attached to the preceding concept 'constriction', in the latter the PP 'in children' will be attached to the head concept 'disease'. Local attachment of PP's (other than those introduced by 'of') only prevails on global attachment (to the head) if certain conceptual conditions are met, such as the nature of the concept types in the PP following a finding concept such as 'constriction'.

2.2.5 Frames

Given a definition of which the head or conceptual type has been established, the parser tries to fill its conceptual template or frame as much as possible. It does so by looking recursively for pre- and postmodifiers (the latter are called subheads), which 'fit' the head (or its modifiers). Fitting here means that the concept type of the governed lexeme corresponds with the concept type of one of the arguments of the template of the governing lexeme. In the 'rheumatoid arthritis' example above e.g. the functional concept type of which 'musculo-skeletal system' is an instantiation, 'fits' or 'fills' the first argument or slot of the concept type rheumatoid arthritis belongs to. M.m. the same can be said for all the other slot-fillers.

From the above it will have become clear that for the representation of conceptual meaning we have chosen for a frame-based system (see e.g. Habel 1985): concept types are defined by frames, i.e. sets of conceptual slots, attributes or features.

3. Usefulness

The parser described here tries to serve a twofold aim. In the first place its aim is practical. By making definitions conceptually explicit it is first of all possible to enhance the access to data bases (by making search items available in a systematic way). Secondly because of the fact that definitional knowledge becomes available in a systematic way, it also becomes possible now to generate from partial conceptual knowledge (answering such questions as: what is the term for the disease caused by HIV?, how is the disease affecting the immune system called ? etc.). Finally, by yielding 'semantically relational knowledge', syntactically ambiguous structures can be more readily solved (think of PP-attachment, cf. he treated the children with epilepsy).

In the second place the system-cum-parser was set up as a pilot project in order to shed some light on such notions as lexicon structure and (power of) concept-oriented parsing. Judging from the results obtained up till now, we dare say that, with regard to the former, a relational-conceptual model of the lexicon offers interesting perspectives (although we still have to tackle in more detail such problems as concept disjunction and non-monotonic default

reasoning), and that, with regard to the parsing, in as far as analytical definitions reflect a conceptual structure, syntactic problems in parsing become by far more feasible to overcome.

4. Bibliography

- Alshawi, H., Analysing the dictionary definitions, in: B.Boguraev and T.Briscoe, (Eds.), Computational Lexicography for Natural Language Processing, London/New York, 1989, 153-169.

- Burkert G. and P.Forster, Representation of semantic knowledge with term subsumption languages, ms., Stuttgart, 1991.

- Habel, C., Das Lexikon in der Forschung der künstlichen Intelligenz, in: C.Schwarze and D.Wunderlich, (Eds.), Handbuch der Lexikologie, Königstein, 1985, 441-474.

- Mars, V. e.a., Eindrapportage Sapiens-project, UT-gedeelte, Twente, 1991.

- Martin, W. e.a., Over Atlex, Relset, Conceptor e.a., VU-bijdrage tot het Sapiens-prototype, Amsterdam, 1991.

- Martin, W. e.a., Dilemma, an automatic lemmatizer, in: Colingua, 1988, 5-62.

- Martin W. and E.ten Pas, Metatools for Terminology, in: Corpusgebaseerde Woordanalyse, Jaarboek 1991, Amsterdam, 1991, 83-99.

- H.Paulussen and W.Martin, Dilemma-2: a lemmatizer-tagger for medical abstracts, Third conference on Applied Natural Language Processing, Trento, 1-3 April 1992, ACL Morristown 1992, 141-146.

- M.Reedijk, Een conceptuele parser voor definities van medische termen, ms., Amsterdam, 1991.

- Winograd, T., Language as a cognitive process. Vol 1: Syntax , Addison-Wesley, 1983.