A robust category guesser for Dutch medical language

Peter Spyns
Katholieke Universiteit Leuven
University Hospital Gasthuisberg
Division of Medical Informatics
Herestraat 49, B-3000 Leuven
peter@uz.kuleuven.ac.be

Abstract

In this paper, we want to describe the architecture and some of the implementation issues of a large scale category guesser for Dutch medical vocabulary. We also provide numerical data on the precision and coverage of this category guesser, which has to cover for the moment only the vocabulary of the cardiology domain. The category guesser uses non-morphologic information (endstring matching) as well as truly morphologic knowledge (inflection, derivation and compounding). Since we deal with a sublanguage some linguistic features are easier to handle (Grishman and Kittredge, 1986), (Sager et al., 1987). Subsequently we will describe in detail the differents parts which interact to successfully identify unknown medical words.

1 Introduction

1.1 NLP in medicine

Medical patient reports consists mainly of free text, combined with results of various laboratories. While numerical data can easily be stored and processed for archiving and research purposes, free text is rather difficult to be treated by computer, although it contains the most relevant information. Several authors put forward the hypothesis that Natural Language Processing (NLP) and Knowledge Representation (KR) of medical discharge summaries have become the key-issues in the domain of intelligent medical information processing (Baud et al., 1992), (Gabrieli and Speth, 1987), (McCray, 1991). However, only a few NLP-driven systems have actually been implemented (Friedman and Johnson, 1992) . For Dutch, a limited prototype has been developed (Spyns, 1991), (Spyns and Adriaens, 1992). A broader system covering a larger part of the Dutch grammar and medical vocabulary is currently under development. This activity forms part of the MENELAS-project [1] . This project comprises a morphological, syntactic, semantic and pragmatic analysis of the medical sublanguage for Dutch, English and French (Spyns et al., 1992). The project also focuses on Knowledge Representation (by means of Conceptual Graphs) (Sowa, 1984), (Volot et al., 1993) and Production Systems (Bouaud and Zweigenbaum, 1992).

1.2 The Category Guesser for Dutch Medical Language

This paper focuses on the morphological and lexical component of the system, which is a combination of a database application and a Prolog rule interpreter. This component is already functioning and is used continuously during the current extension of the coverage of the Dutch grammar (Spyns et al., 1993). The importance of morphologic analysis of medical vocabulary has been widely recognised (Wingert, 1985), (Wolff, 1984), (Dujols et al., 1991), (Pacak and Pratt, 1969) (Pacak and Pratt, 1978), (Norton, 1983).

In the following sections, we will describe the different parts which interact to identify the word forms of a given sentence. The various stages of the analysis of the word forms are described. A major distinction can be made between forms "known by the system" (= stored in the dictionary cf. section 2) and unknown forms whose linguistic characteristics need to be computed and are hypothetical. The latter can be based on morphologic knowledge (section 3) or other heuristics (sections 4, 5 & 6). Each section is illustrated by an example or some implementation details. A schematic overview of the architecture of the category guesser is presented in section 7. The subsequent section (8) is devoted to the evaluation, which will guide the further elaboration of the here described category guesser. The paper ends with a conclusion and discussion (section 9).

[1] The MENELAS-project (AIM #2023) is financed by the Directorate General XIII of the European Community (Zweigenbaum and others, 1991).

```
[lex:geprobeerd,nl_lu:geprobeerd,cat:n,nb:sing,pers:3]
[lex:geprobeerd,nl_lu:geprobeerd,cat:adj,adjtype:ord,adj_e:no]
[lex:geprobeerd,nl_lu:proberen,cat:v,pers:nil,nb:nil,tense:nil,vform:pastpart]
[lex:geprobeerd,nl_lu:proberen,cat:adj,adjtype:papa,adj_e:no]
[lex:geprobeerd,nl_lu:geprobeerden,cat:v,pers:nil,nb:nil,tense:nil,vform:pastpart]
[lex:geprobeerd,nl_lu:geprobeerden,cat:adj,adjtype:papa,adj_e:no]
[lex:geprobeerd,nl_lu:geproberen,cat:adj,adjtype:papa,adj_e:no]
[lex:geprobeerd,nl_lu:geprobeerden,cat:v,pers:1,nb:sing,tense:pres,vform:finite]
```

Figure 1: Example of Cohort for "geprobeerd"

2 Full Form Dictionary

The lexical database for Dutch was built using several resources: an existing electronic valency dictionary and a list of words extracted from a medical corpus (cardiology patient discharge summaries). The already existing electronic dictionary (resulting from the K.U. Leuven PROTON-project (Dehaspe and Van Langendonck,) and the newly coded entries were converted and merged into a common representation in a relational database (Dehaspe, 1993).

It is intended to use the category guesser (cf. infra) as little as possible. To that extent, the dictionary is conceived as a full-form dictionary. Currently, there are some 100.000 full forms in the lexical database (which is some 8000 non inflected forms). However, since an exhaustive dictionary is an unrealistic assumption, a category guesser handles all the unknown word forms.

The unknown words trigger a set of rules to identify the surface form, to attribute syntactic categories to it, and to calculate the possible canonical form(s). The category guesser can also enhance the robustness of the larger NLP-system since misspelled words can receive, to a certain extent, correct syntactic features. To reach this aim, the category guesser combines morphologic (3) as well as non morphologic knowledge (sections 4 & 6).

3 Morphological Analysis

3.1 Preliminary Remarks

The morphological analyser consists mainly of three sections, which correspond more or less to the three linguistic operations on words: inflection, derivation and compounding. However, from an implementational point of view, the boundaries between derivation and compounding are defined in a different way. The compounds, created by agglutination or combined by means of a hyphen are computationally treated as non-compounds. This implies that the same segmentation routine can be used for the computation of derivations and monolithical compounds (Spyns and De Wachter, 1995).

3.2 Inflection

The inflection analyser produces one or more bundles of morphosyntactic feature value pairs for each submitted surface form (= cohort). The generated feature bundles comprise, among other features, the surface form (lex), the supposed canonical form (nl_lu) as well as its category (cat) [2] . A reduced example of the cohort produced for "geprobeerd" (Eng.: "tried") follows (see figure 1).

The initial cohort will later on be reduced as much as possible (the ideal result in most cases being a single feature bundle). Therefore, a cascading priority system has been defined. The attribute "morf" expresses the quality of the analysis, possible values being segm, suffix, string or guess with segm > suffix > string > guess. More details on this will be given below.

Only the feature bundles of supposed nouns, verbs, adjectives and adverbs (i.e. the open categories) are admitted in the initial set of hazardous analyses or cohort.

3.3 Segmentation

Derivation and monolithical compounding are used to try and identify as many as possible of the canonical forms computed by the inflectional analyser. The starting principle here is that the right part of the computed canonical form usually constitutes the grammatical head of the whole word. The whole word thus inherits the feature-bundle associated with its right part (Selkirk, 1982, p.150) [3] .

In opposition to William (Williams, 1983) & Selkirk (Selkirk, 1982), we do not allow inflectional suffixes to be heads. The right part can be found in the dictionary (monolithical compounding) or in a list of suffixes (derivation). In the current segmentation program, the major part of this list contains medical suffixes, which constitute a clearly definable

[2] v for verbs, adj for adjectives and n for noun; others are nb [sing or plur] for number, pers [1, 2, 3 or nil] for person.

[3] We are fully aware that linguistic reality is more complex: e.g. some derivations (f.i. Dutch diminutifs cf. (Ritchie et al., 1992)) are regarded as left headed. Maybe they should be treated computationally by the inflectional analyser.

set that is fairly regular in its (morphological and syntactic) behaviour (Dujols et al., 1991). Below (see figure 2) one can find an extract of the suffix list.

```
suffix([a,r,i,s],[cat:adj,nb:sing]).
suffix([a,a,l],[cat:adj,nb:sing]).
suffix([i,e],[cat:n,nb:sing]).
```

Figure 2: Examples of Suffixes with Feature Bundle

The computed canonical form is scanned and segmentated from right to left. All possible solutions are generated by a failure driven loop (no exclusive longest match principle). The segmentation routine which tries to identify a right part (head:dict or head:suffix) and then tries to recognize the remaining left part. If this succeeds, the segmentation is complete (morf:segm). Otherwise, it is only partial (morf: suffix).

At the moment, only noun noun compounds are treated. Many medical noun noun compounds combine a medical non head part with a non medical head part (f.i. hartziekte - Eng.: heartdisease).

Only those feature bundles of the cohort are kept that are compatible (by means of graph-unification) with the feature bundle associated with the head part (suffix or dictionary entry). At this stage of filtering, the feature cat (syntactic category) plays a most prominent role.

4 Endstring Matching

When nothing can be predicted by means of morphology, another heuristic will be applied to reduce the set of remaining possible morphological analyses. This stage will focus more on the general language words. It is based on a series of endstrings (not limited by morphological boundaries) which determine the category of a word. Only the open syntactic classes are taken into account (noun, verb, adjective and adverb). Some endstrings uniquely identify the category of a word while others are more equivocal. The latter are correlated with two or even three categories. The necessary linguistic knowledge to build a list of non-inflected endstrings and their associated category (or categories) was found in Lemmens (Lemmens, 1989). Some combinations of an endstring and its category are shown below (see figure 3).

```
end([d,r,e,e|_],[v,adj],[eerd]).
end([l,e,e,i|_],[adj,n],[ieel]).
end([l,e,i|_],[adj],[iel]).
end([e,m,s,i|_],[n],[isme]).
```

Figure 3: Some endstring-category combinations

When a computed lexical form is presented to the endstring matcher, the above mentioned list is checked to see if an endstring constitutes the endpart of the submitted word. In fact, the surface form as well as the hypothetical canonical form of the feature bundle are submitted to the endstring matcher. Only the categories resulting of both matching processes (= the intersection) are finally retained. Subsequently, the feature bundle(s) of the cohort containing the proposed syntactic category are extended with an extra featurevaluepair (morf:string). Below (see figure 4) the result of endstring matching applied to the verb "geprobeerd" (Eng.: "tried") is shown (rule with ending -eerd applies) [4] .

The inflection rules were able to produce a canonical form together with its category which the endstring matcher considers correct. This implies that the inflection rule was correctly triggered and applied. As a corollary, the other syntactic information in such a validated feature bundle (with morf:string) is supposed to be correct as well. However, many syntactic features are underspecified [5] .

5 Default or Catch All Rule

If none of the aforementioned cases apply, the computed canonical forms and its corresponding grammatical features are pure guesses. The complete cohort is retained and each of its feature bundles is extended with one extra feature morf: guess.

6 Final selection of the set of solutions

After the stages mentioned above, only a subset of feature bundles of the cohort will contain the feature morf. All of these feature bundles contain morphosyntactic information that is validated by the mentioned heuristics (cf. supra) [6] . This subset is retained and passed to the syntactic parser. When segmentations of both types (complete versus partial) are produced, the latter (morf:suffix) are discarded in favour of the former (morf:segm). In that case, endstring matching nor the catch all rule is applied.

7 Schematic Overview

Below, one can find a more formal description and a schematic overview (see figure 5) of the category guesser.

[4] We assume of course that the verb does not appear in the dictionaries.

[5] The syntactic analyser requires the presence of some linguistic features — even when underspecified — in the feature bundle.

[6] When the default rule applied, the subset will be identical to the complete cohort. Validation is a too strong word in this case.

```
[lex:geprobeerd,cat:adj,nl_lu:proberen,adjtype:papa,adj_e:no,morf:string]
[lex:geprobeerd,cat:adj,nl_lu:geproberen,adjtype:papa,adj_e:no,morf:string]
[lex:geprobeerd,cat:v,nl_lu:proberen,vform:pastpart,vtype:main,morf:string]
[lex:geprobeerd,cat:v,nl_lu:geproberen,vform:pastpart,vtype:main,morf:string]
```

Figure 4: Endstring Matching Applied to "geprobeerd"

Input: [lex:W] the unknown surface form W
Data : FDAG: the linguistic information (feature bundle) associated with F
FDAG = [nl_lu:W', cat:x, nb:y, ...]
DDAG: the dictionary entry of a canonical form W'
DDAG = [nl_lu:W', cat:x, frame:z, ...]
EDAG: the linguistic information (feature bundle) associated with an endmorpheme
EDAG = [cat:x, head:dict, ...] or EDAG = [cat:x, head:suffix, ...]
RDAG: the category provided by the endstringmatcher
RDAG = [cat:x]

Function: $\mathcal{F}$: maps W to a hypothetical canonical form W$'$ (inflection rule)
$\mathcal{U}$: unifies two feature bundles
$\mathcal{S}_1$: segmentates W $'$ in a left part (L) and endmorpheme (E)
$\mathcal{S}_2$: segmentates L in several morphemes ($L_1, \ldots, L_n$)

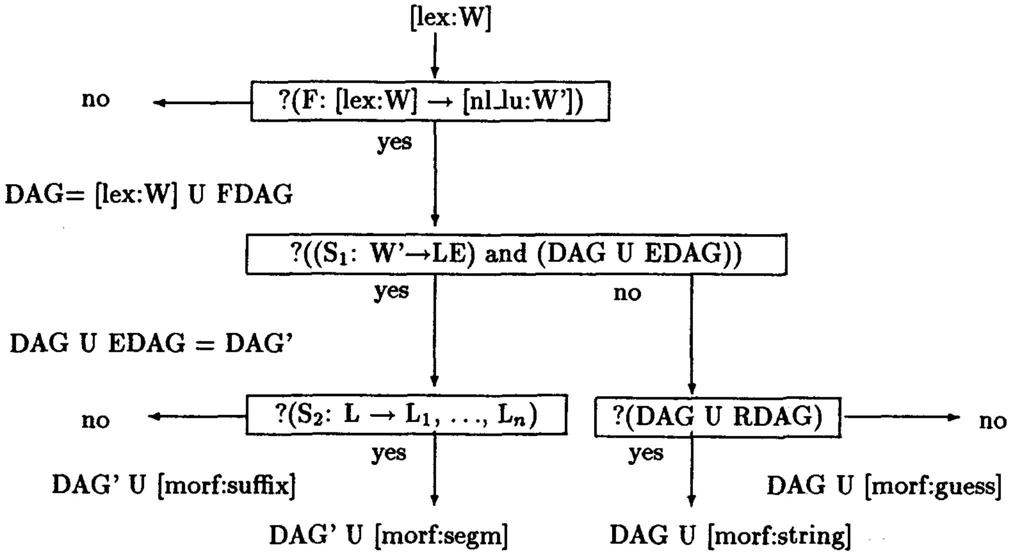

Figure 5: The High Level Algorithm

8 Some Results and Statistics

To examine the effectiveness of the category guesser, all the words from the corpus not appearing in the dictionary were submitted to the analyser. The total number of unknown words was 2832. Manual categorisation revealed the presence of 679 adjectives, 2056 nouns, 82 verbs. The 2832 unique unknown forms lead to the generation of 6342 supposed analyses, which means that for every unknown form 2.4 possible canonical forms are retained. We consider the case when an unknown surface form receives more than two different categories as a guess. Guesses are always interpreted as bad. If the category guesser is not able to attribute a correct category, the result is regarded as bad. Once a correct category, even concurrently with an incorrect one, is assigned to the submitted word, the outcome is perceived as good. [7] As the main concern lies with the syntactic characteristics, we did not consider an erroneously calculated canonical form as a reason to reject the complete feature bundle. Manual examination of the results permits us to state that 83.4 % of the unknown forms are correctly identified. We consider the result as fairly good and are convinced that refinements can lead to an even better result. The linguistic coverage can be still be improved by adding rules in order to treat comparatives and superlatives.

9 Discussion

A choice was to be made between keeping more potential analyses likely to be correct versus restricting the cohort to one (or a limited set of) analysis which may be incorrect. As a general strategy, we prefer to restrict as early as possible the search space on all the levels of the language understanding system. Otherwise useless hypotheses will be propagated through the whole system causing a combinatorial explosion. However, this attitude can lead to the rejection of valid solutions and, in the worst case, can be responsible for a complete failure of the language understanding system.

A possible optimisation resides in the storage of the medical suffixes and endstrings in their inflected forms. They could be integrated in the already existing full form dictionary. In order to accelerate the decomposition phases, the morphemes or strings could be stored in reversed order.

These reorganisations of the data structures also influence the high level algorithm (cf. section 7). Since all the words, suffixes and endstrings would be stored in the database as full forms, the inflectional analyser (cf. section 3.2) would be merely needed for the computation of a hypothetical canonical form and its syntactic characteristics when applying the catch all rule. This leads without any doubt to a faster execution of the category guesser as a whole. As a corollary, the overall architecture of the entire component becomes simpler and more homogeneous.

[7] Sometimes the surface form alone does not permit an unequivocal categorization (f.i. in principle, a Dutch noun formally equals the first person singular present of a regular verb).

References

R.H. Baud, A.-M. Rassinoux, and J.R. Scherrer. 1992. Natural language processing and semantical representation of medical texts. *Methods of Information in Medicine*, 31:117 - 125.

J. Bouaud and P. Zweigenbaum. 1992. A reconstruction of conceptual graphs on top of a production system. In *Proceedings of the 7th Annual Workshop on Conceptual Graphs*, Las Cruces.

L. Dehaspe and W. Van Langendonck. Automated valency dictionary of Dutch verbs. to appear.

L. Dehaspe. 1993. Menelas report on the building of the lexical database. Technical Report 93-002, K.U. Leuven - Dept. of Medical Informatics.

P. Dujols, P. Aubas, C. Baylon, and F. Grémy. 1991. Morphosemantic analysis and translation of medical compound terms. *Methods of Information in Medicine*, 30:30 - 35.

C. Friedman and S.B. Johnson. 1992. Medical text processing: Past achievements, future directions. In M.J. Ball and M.F. Collen, editors, *Aspects of the Computer-based Patient Record*, pages 212 - 228. Springer - Verlag, Berlin.

E.R. Gabrieli and D.J. Speth. 1987. Computer processing of discharge summaries. In *Proceedings of SCAMC 87*, pages 137 - 140.

R. Grishman and R. Kittredge. 1986. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

M. Lemmens. 1989. A critical study of the defaulters in the Belgian METAL-system ©, and a design of a morphologically guided category guesser. Master's thesis, K.U.Leuven. [in Dutch].

A.T. McCray. 1991. Natural language processing for intelligent information retrieval. In J.H. Nagel and W.M. Smith, editors, *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 1160 - 1161, Orlando.

M.G. Norton, L.M. & Pacak. 1983. Morphosemantic analysis of compound word forms denoting surgical procedures. *Methods of Information in Medicine*, 22(1):29 - 36.

M.G. Pacak and A.W. Pratt. 1969. Identification and transformation of terminal morphemes in medical English (part 1). *Methods of Information in Medicine*, 8(2):84 – 90.

M.G. Pacak and A.W. Pratt. 1978. Identification and transformation of terminal morphemes in medical English (part 2). *Methods of Information in Medicine*, 17(2):95 – 100.

G.D. Ritchie, G.J. Russell, A.W. Black, and S.G. Pulman. 1992. *Computational Morphology: Practical Mechanisms for the English Lexicon*. MIT Press.

N. Sager, C. Friedman, and M. Lyman. 1987. *Medical Language Processing: Computer Management of Narrative Data*. Addison Wesley, Reading, Massachussets.

E. Selkirk. 1982. *The Syntax of Words*. MIT Press.

J.F. Sowa. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, London.

P. Spyns and G. Adriaens. 1992. Applying and improving the restriction grammar approach for Dutch patient discharge summaries. In *Proceedings of COLING 92*, pages 1164 – 1168.

P. Spyns and L. De Wachter. 1995. Morphological analysis of Dutch medical compounds and derivations. *ITL Review of Applied Linguistics*. to appear.

P. Spyns, P. Zweigenbaum, and J.L. Willems. 1992. Representation and extraction of information from patient discharge summaries by means of natural langage processing. In *Proceedings of MIC92*, Rotterdam. [in Dutch].

P. Spyns, L. Dehaspe, and J.L. Willems. 1993. The Menelas syntactic analysis component for Dutch. Delivrable report AIM-Menelas #6.

P. Spyns. 1991. A prototype of a semi-automated encoder for medical discharge summaries. Master's thesis, K.U. Leuven. [in Dutch].

F. Volot, P. Zweigenbaum, B. Bachimont, M. Ben Said, J. M., Bouaud, M. Fieschi, and J.F. Boisvieux. 1993. Structuration and acquisition of medical knowledge (using UMLS in the conceptual graph formalism). In *Proceedings of SCAMC 93*.

E. Williams. 1983. On the notions 'lexically related' and 'head of a word'. *Linguistic Inquiry*.

F. Wingert. 1985. Morphologic analysis of compound words. *Methods of Information in Medicine*, 24:155 – 162.

S. Wolff. 1984. The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. *Methods of Information in Medicine*, 23:195 – 203.

P. Zweigenbaum et al. 1991. AIM project #2023 technical annex: An access system for medical records using natural language. manuscript.