DILEMMA-2: A LEMMATIZER-TAGGER FOR MEDICAL ABSTRACTS

Hans Paulussen
Facultés Universitaires Notre-Dame de la Paix,
rue de Bruxelles 61, B-5000 Namur, Belgium
phone +32-81-72.41.37, fax +32-81-23.03.91, e-mail hpaulus@cc.fundp.ac.be

Willy Martin
Vrije Universiteit, De Boelelaan 1105, NL-1007 MC Amsterdam, The Netherlands
phone +31-20-548.37.63, fax +31-20-661.30.54, e-mail lexico@let.vu.nl

Abstract

This paper reports on the development of DILEMMA-2*, a lemmatizer-tagger for the *sublanguage* of *medical abstracts*. The program is an extension of DILEMMA-1, a lemmatizer-tagger for *general* English texts.

In the first section a brief outline is given of DILEMMA-1. Particular attention is paid to the original concept of a *default category* which is linked with a *categorial graph* by means of a pointer system. In the second section we show why DILEMMA-1 was not able to get a suitable score when lemmatizing medical abstracts, the main reason being the inability to recognize *sublanguage specific vocabulary*. In the next section a description is given of the most important errors along with their solutions; these errors are then categorized as *gaps* or *wrong assignments*. The former could be dealt with in either a *suffix list* or a *gaps filler default*. The latter mainly concerned wrongly assigned past participles and errors on noun, verb or adjective assignment.

After implementation of the proposed solutions, a comparison is made between the results of DILEMMA-1 and DILEMMA-2, showing that the results of DILEMMA-1 have been improved substantially within a sublanguage context, and this by using linguistic, i.e. sublanguage, knowledge, thus avoiding ad hoc remedies.

0 Introduction

In this paper we describe DILEMMA-2, a lemmatizer-tagger for medical abstracts, which is an updated version of DILEMMA-1, a lemmatizer-tagger for general texts. After a brief outline of DILEMMA-1 we give a description of the types of errors we found when running the general lemmatizer on medical abstracts. This is followed by some examples of the solutions we proposed and implemented into DILEMMA-2. Finally, the results of DILEMMA-1 and DILEMMA-2 are compared, showing that a sublanguage approach can lead to workable results in the development of real world applications.

1 DILEMMA-1

DILEMMA-1 is an automatic lemmatizer-tagger for general English texts, developed at the University of Antwerp during the academic year 1985-1986 (see [MARTIN 88b]). For each word of the text it tries to find its lemma (or dictionary entry form) and its grammatical category, and subcategories (or specifiers) where necessary.

Being a lemmatizer, not a parser, DILEMMA-1 is as such limited to a relatively basic level of syntactic analysis, which however can be used as input to a more powerful syntactic analyzer. In this way, a lemmatizer can be considered an invaluable tool for corpus linguistics.

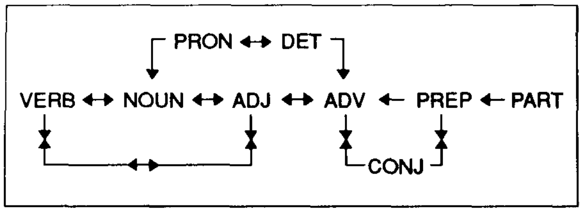
To carry out the task of assigning grammar categories and possible specifiers DILEMMA-1 looks at word forms from four different points-of-view. First of all word forms are looked at out-of-context (dictionary look-up, morphological procedures). In a second step the immediate context is taken into account: word forms are analyzed and checked by looking at the words immediately preceding and following them. In a third step, the proto-syntactic module, a larger context (such as verb patterns) is taken into account. Finally, in the temporary memory, word forms are checked by looking at the whole text.

Like most modern lemmatizers, DILEMMA-1 uses as much linguistic knowledge as possible, by translating any *regularity* on the lexico-morphological level into

* DILEMMA-2 was developed as part of a research contract for *Elsevier Science Publishers* (ESP), Amsterdam, The Netherlands. The development of DILEMMA-1was carried out as part of contract research for *Van Dale Lexicografie* Publishers, Utrecht, The Netherlands.

rules, thus keeping the dictionary small. But in the case of DILEMMA-1 the size of the dictionary is extremely small when compared to other lemmatizer-taggers: a little over 3600 words. For a comparison between DILEMMA-1 and CLAWS, another well-known, lemmatizer-tagger for English, see [MARTIN 88b]. In passing it can be noted that DILEMMA-1 uses a dictionary only half the size of that used in CLAWS. This smallness is due to criteria adopted on the macro- and micro-level of the dictionary. The vertical *macro-level* is concerned with the words to select as entry, whereas the horizontal *micro-level* deals with the information to store next to the dictionary entry.

Figure 1: successor and predecessor relations between the categories



On the macro-level, the dictionary entries are selected according to the following three principles: *frequency*, *closure* (of classes, e.g. prepositions) and *irregularity* (e.g. irregular verbs, irregular plurals). In other words, the dictionary only contains words which are either frequent or which belong to closed classes or which cannot be deduced from the grammar of the English language. To construct the list of frequent words, we used *Van Dale's English-Dutch Dictionary* ([MARTIN 89]), where very frequent words are labeled F4 or F3. These frequency codes were the result of an earlier research project ([MARTIN 83], [MARTIN 88a]).

On the micro-level, categorial information is stored *preferentially*. Each word is given a preferential default category which can shift to other categories along a categorial graph manipulated by the program (see Fig. 1). This way of storing categorial information is based on the fact that DILEMMA-1 also tries to look for regularity in the categories words can have, and that is what makes it so different from other lemmatizers.

Table 1: part of the DILEMMA dictionary

wordform	lemma	DC	ptr	specifiers
king	king	noun	l	
kiss	kiss	verb	r	
kit	kit	noun	l	
kitchen	kitchen	noun	n	
knee	knee	noun	l	
kneel	kneel	verb	n	
knelt	kneel	verb	n	pastpapa
knew	know	verb	n	past

Being a morphologically poor language, English has a large number of grammatical homonyms. DILEMMA-1 starts from the assumption (i) that English words can have different categories, (ii) that each word has a default category (DC), and (iii) that the necessary categorial shifts can be systematized. The DC, which is the main category of a word, is established on the basis of *frequency, analogy* and/or *meaning*. Next to a DC, each word in the dictionary (see Table 1) has a pointer (left, right or neither) indicating the direction in which a category can shift through a *categorial graph* which was established after calculating the combination and frequency of categories. The word "kiss", for example, has the category 'verb' as DC, and can shift 'right' to the 'noun' category. All categorial shifts are guided by condition-action rules in the rule component of the DILEMMA-1-program. Note also that the categories 'numeral' and 'interjection' are not integrated in the graph. The numeral only has a predecessor, viz. noun; the interjection has neither a predecessor nor a successor.

Table 2: DILEMMA-1 output sample

```
   The           the          det  art
   inclination   inclination  noun sg
   here          here         adv
   is            be           verb pres 3 sg
   to            to           part
   accept        accept       verb inf
   a             a            det  art
** de
** facto
   cease-fire    cease-fire   noun sg
   in            in           prep
   Laos          Laos         noun prop
   ,
   rather        rather       adv
   than          than         conj
   continue      continue     verb
   to            to           part
   insist        insist       verb inf
   on            on           prep
   a             a            det  art
   verifi-       verifi-      noun sg
      cation        cation
   of            of           prep
   the           the          det  art
   cease-fire    cease-fire   noun sg
   by            by           prep
   the           the          det  art
   interna-      interna-     adj
      tional        tional
   control       control      noun sg
   commission    commission   noun sg
   before        before       prep
   partici-      partici-     verb ing
      pating        pating
   in            in           prep
   the           the          det  art
   Geneva        Geneva       noun prop
   conference    conference   noun sg
   .
```

The use of categorial information and pointers changes the dictionary into an economical and *dynamic* set of lexemes. This is maybe the most striking feature of the modular architecture of DILEMMA-1, and it explains also why the program can run even within a PC-environment. For a fuller account of the DILEMMA-1-program we refer to [MARTIN 88b].

An output sample of DILEMMA-1 is shown in Table 2, which is a sentence from the BROWN-corpus (see [KUCERA 67]). The first column is the text, the second is the lemmatized form, and the following columns give the category and possible specifiers. When a word is not recognized, or when recognition is doubtful, it is flagged by a double asterisk.

DILEMMA-1 was tested on a number of general language text samples and proved to be a very powerful tool. A sample of error analysis on 6 texts taken from a standard British English corpus (the LOB corpus [JOHANSSON 78]) shows e.g. that for general language texts, DILEMMA-1's success rate does not drop below 90%, nor does it exceed 97%, on the average leading to a score of ± 93.50% (see Table 3).

Table 3: results LOB samples

text number	total of word-tokens	total of errors made	error percentage	relative success rate
1	123	6	4.87	95.13
2	183	7	3.82	96.18
3	196	7	3.57	96.43
4	143	14	9.79	90.21
5	239	17	7.11	92.89
6	246	19	7.72	92.28
1-6	1130	70	6.14	93.86

Nevertheless, when DILEMMA-1 was tested on a number of medical abstracts, its *scoring reference point* of 93.50% was not reached at all. 'Best results' were more likely to lie within the 90% area, the average being about 86% (see Table 6). The object of this research project was how to bring back the success rate for lemmatizing medical abstracts, without changing the philosophy behind the DILEMMA-1-program, which is developed as a robust, preferential, dynamic system in which items can take different values governed by constraints. Moreover, in a language such as English, categories are often functional instead of lexical (which explains, in part, the small size of the lexicon).

2 A Sublanguage Approach

When running the DILEMMA-1 program on medical abstracts, we found that most errors are related to the *sublanguage* of *medical abstracts*. For example, most gaps in the output are due to a lack of *sublanguage specific vocabulary* in the DILEMMA-1 dictionary: e.g. *astrocyte, fibrillary, acidic, GFAP*. Another point which supports the idea of sublanguage influence is that the more abstracts resemble general language texts, the more their results lie within the general language area. In an extreme case there was only one error in a text of 42 words (success rate = 97.62%). Very unlike the average medical abstract, this text showed no symbols or abbreviations, and it had short, non-complex sentence structures. For a fuller account of lexical differences between sublanguage and general language lexicons, see [MCNAUGHT 91].

An example showing that the sublanguage features are not solely confined to the lexical level is the following sentence, where 'counts' — which can be either 'verb' or 'noun' — must shift from 'verb' to 'noun' when found at the beginning of a sentence:

e.g. *Counts* of neocortical cells did not reveal differences in cell numbers.

This categorial shift is a sublanguage shift, as categorial and syntactic ambiguity does not exist here, i.e. sentence initial verbal constructions such as imperatives and questions do not occur in the sublanguage of medical abstracts.

To improve the DILEMMA-1 program, we not only tried to tackle the problems from a sublanguage approach, but we also decided to implement all program adaptations in a separate module which can be called up by the user whenever he wants. Such a modular architecture makes it easier to adapt the program to another sublanguage.

Although not new, the sublanguage approach is being adopted more and more in the implementation of real world applications, where computational linguists are constantly confronted with how to organize the vast amount of world knowledge. The domains can be very diverse as can be seen in the examples of [CHEVALIER 78] (automatic translation of weather forecasts), [DEVILLE 89] (automatic man-machine dialogue system handling requests for administrative information) and [PALMER 90] (physics world problems for college students involving pulley systems). Only by strictly defining the limits of the application domain can one write programs without having to resort to brute force techniques.

Even if we stress the sublinguistic character of the errors in the DILEMMA-1 output, there were of course also a number of general errors, most of which could not be solved within the context of the DILEMMA-1 framework which presupposes no (clause) syntactic knowledge. In the rest of this paper we will focus our attention on the modifications added at the sublanguage level.

3 DILEMMA-2

DILEMMA-2 is the result of the corrections we made to DILEMMA-1 within the context of medical abstracts. The type of errors encountered were either gaps or wrong assignments (see Table 6 at the end of this section).

3.1 Gaps

As explained in section 2, most gaps were sublanguage specific words. Putting the missing scientific terms in the dictionary was not considered a good solution: this would have been against the basic principle of the DILEMMA concept, which was to keep the dictionary as small as possible; in any case, it would have been a practically impossible task (ESP has a database of more than 100.000 scientific terms). In as far as the missing terms showed some *regularity* at the lexico-morphological and syntactic level, they could be dealt with outside the dictionary, by using a *sublanguage specific suffix list* and a *sublanguage specific gaps filler default*. The former has been arrived at by considering medical sublanguage from a broad point-of-view, situating it within the functional domain of scientific writing. Table 4 gives a sample of scientific suffixes which have categorial power and which are typical for the formation of medical terms in a broad sense of the term.

Table 4: part of the scientific suffix list

emia	noun	n
enchyma	noun	n
esis	noun	n
escent	adj	n
ferous	adj	n
fuge	noun	r
gen	noun	l
gene	noun	n

As to the default for the remaining gaps, it became apparent from the sample analysis that medical texts, like most scientific texts, *heavily nominalize*. As a result, we expect remaining gaps to be (part of) NPs. Given the contents of the existing DILEMMA dictionary this leads us to a choice between (predominantly) nouns and adjectives. Therefore, only in a last module is DILEMMA-2 allowed to fill out all remaining gaps as *nouns* unless these gaps:

(a) occur in prototypical patterns for adjectives such as (Pron is mentioned here, because it has not been shifted to Det):

```
Det X Noun
Pron X Noun
```

(b) occur in prototypical patterns for verbs, such as possible candidates for a Verb+Object NP:

```
X Adj (Noun)
   Det (Adj) (Noun)
```

(c) end in -al, -ar, -ile, -ine, -y; unless followed by Noun, then they are shifted to Adjective. These are typical endings which can also yield adjectives and/or verbs.

In these cases the gaps remain unfilled and are flagged for further processing by a higher module, such as a clause module or a syntactic parser.

3.2 Wrong Assignments

Although wrong assignments are far less frequent than gaps, they can be important in so far as they can give rise to wrong results in further processing (e.g. in establishing NPs), and in so far as they are no longer easily recoverable. The most important of these errors are related to either those cases where specification of *simple past* or *past participle* is difficult to distinguish, or to errors concerning *noun*, *verb* or *adjective* assignment. Another problem, we will not deal with in this paper, is the ample use of differently structured *abbreviations*, such as: *MH, VAHR*, *b.i.d.*, *mRna*. These were handled by an abbreviations procedure.

A major problem, well known in English tagging and so not only restricted to medical texts, is the wrong specification of verbs which can be either *simple past* or *past participle*. As long as the specifier is not disambiguated, it is referred to as PAST_PAPA. In the following example both 'revealed' and 'increased' were assigned PAST, whereas 'increased' should have been assigned PAPA:

e.g. Western blot analysis *revealed increased* levels of GFAP in Mo(br/y) forebrain and cerebellum.

In the context of another ESP project on automatic indexing of medical abstracts, it is important to correctly delineate NPs and therefore to recognize PAPA's. However, within the framework of a lemmatizer-tagger this is not an easy task. Again, a sublanguage approach is of great help here. Attributive PAPA's in our sublanguage occur much more often than in general English texts. Consequently, in some very strictly defined contexts we could partly disambiguate the PAST_PAPA problem, as in the following three examples:

(a) When a PAST_PAPA is preceded by an ING-form and followed by a noun, select PAPA: e.g. physicians are *expressing increased* willingness.

(b) When a PAST_PAPA is preceded by a noun and followed by a preposition or a particle, select PAPA: e.g. *cells isolated* from ...

(c) When a PAST_PAPA is found at the beginning of a sentence, select PAPA: e.g. *Affected* males suffer profound deficits in oxidative metabolism.

Each case was implemented in a condition-action rule, so that example (a), when written as a C-function looks as follows:

```
if ((thisspec == C_past)
   && (this[0].spec2 == C_papa)
   && (prevspec == C_ing)
   && (nextcat == C_noun))
{
   thisspec = C_papa;
   this[0].spec2 = C_nullspec;
}
```

This function states that

```
if
     the specifier of the selected
     word is C_past
 and it has an alternative
     specifier C_papa
 and the specifier of the preceding
     word is C_ing
 and the category of the next word
     is C_noun
then
     change the specifier of the
     selected word into C_papa
 and eliminate the alternative specifier
```

As stated above, our proposal of PAST_PAPA rules was based on observations of the Elsevier corpus of medical abstracts at our disposal. We found that the contexts in which wrongly coded attributive PAPA's occur, can — as a rule — be characterized as:

(a) object NPs following a verb (e.g. and have found *marked* changes);

(b) NPs not (directly) following a verb (e.g. once in individual association and once in a combined, *fixed* preparation);

(c) complex NPs (e.g. after a co-ordinating conjunction);

(d) subject NPs at the beginning of sentence (e.g. *Affected* males suffer profound deficits ...).

In the case of errors concerning noun, verb or adjective assignment, we encountered similar context problems, and again, only in very strict contexts could a rule be applied.

3.3 Results

After implementation of the proposed solutions, we compared the success rates of DILEMMA-1 and DILEMMA-2. An output sample, based on six randomly selected texts from 30 ESP samples, is given in Table 5. The results, summarized in Table 6, show that the modifications in DILEMMA-2 have improved the success rate considerably, even passing the scoring reference point of DILEMMA-1 (93.5% vs. 96%).

Table 5: part of lemmatized medical abstract

A	a	det	art
comparative	comparative	adj	
bioavailability	bioavailability	noun	sg
study	study	noun	sg
of	of	prep	
the	the	det	art
antituberculosis	antituberculosis	noun	sg
drugs	drug	noun	pl
isoniazid	isoniazid	noun	sg
,			
rifampin	rifampin	noun	sg
,			
and	and	conj	coor
pyrazinamide	pyrazinamide	noun	sg
was	be	verb	past
carried	carry	verb	papa
out	out	adv	
in	in	prep	
a	a	det	art
group	group	noun	sg
of	of	prep	
10	10	num	
healthy	healthy	adj	
volunteers	volunteer	noun	pl
after	after	prep	
administration	administration	noun	sg
of	of	prep	
the	the	det	art
three	three	num	
compounds	compound	noun	pl
,			
once	once	adv	
in	in	prep	
individual	individual	adj	
association	association	noun	sg
and	and	conj	coor
once	once	adv	
in	in	prep	
a	a	det	art
combined	combined	adj	
,			
fixed	fix	verb	papa
preparation	preparation	noun	sg
.			

4 Conclusions

In this paper we have shown that it is possible to adapt a general lemmatizer-tagger for the specific purpose of lemmatizing medical abstracts, and this by using linguistic knowledge, rather than any ad hoc solutions.

A dynamic, preferential, and constraint-based system with a very small dictionary such as DILEMMA-1 lends itself particularly well to such an adaptation, as no massive lexical importation is necessary, although it looks like that at first sight. Instead, mainly morphological and syntactic sublanguage knowledge has been made use of, leading to a *sublanguage suffix list*, a *sublanguage default procedure* and improved functions of

PAST_PAPA specifiers. The program was modified modularly so that adaptations to other sublanguages could easily be carried out. In this way DILEMMA-2 is at present used as an invaluable pre-processor for an abstracts indexing program at *Elsevier Science Publishers*, Amsterdam. It should be clear that, generally speaking, lemmatizing-tagging performance correlates in a positive way with indexing and information retrieval in that it paves the way for finding NPs (as possible index terms) and that it abstracts away from inflectional variation (thus covering morphological variants in search words).

Although some minor corrections are still possible (for example in the lemmatizing procedures), we believe that DILEMMA-1 cannot reach a higher score within the limits it is conceived for. Interestingly, improvements are possible 'horizontally', thus allowing extensions for other sublanguages; further 'vertical' refinements should be left to more powerful tools. On the other hand, by bringing in more flags to signal uncertainties, DILEMMA-2 could reach, with some small guided post-editing effort, a near-perfect score.

Table 6: summary dilemmatized ESP abstracts

	text 1	text 2	text 3	text 4	text 5	text 6	TOTAL	success rate
total amount of words	186	86	150	176	242	233	1073	
wrong assignments	12	3	6	5	22	5	53	
gaps	21	7	14	15	32	11	100	
total amount of errors	33	10	20	20	54	16	153 (14.25%)	DILEMMA-1 85.75 %
total amount of corrections	25	7	16	14	37	9	108	
remaining errors	8	3	4	6	17	7	45 (4.19%)	DILEMMA-2 95.81 %

References

[CHEVALIER 78]
Chevalier, M.: *TAUM-METEO: description du système*, Groupe de recherche en traduction automatique, University of Montreal, 1978.

[DEVILLE 89]
Deville, G.: *Modelization of task-oriented utterances in a man-machine dialogue system*, PhD dissertation, UIA: Antwerpen, 1989.

[JOHANSSON 78]
Johansson, S., G. Leech & H. Goodluck (eds.): *Manual of information to accompany the Lancaster-Oslo-Bergen corpus of British English, for use with digital computers*, University of Oslo, 1978.

[KUCERA 67]
Kucera, H. & W. N. Francis (eds.): *Computational analysis of present-day American English*, Providence, Rhode Island, 1967.

[MARTIN 83]
Martin, W.: "The construction of a basic vocabulary: an objective-subjective approach" in *Linguistica Computazionale III*, 183-197, 1983.

[MARTIN 88a]
Martin, W.: "Variation in lexical frequency" in P. van Reenen & K. van Reenen-Stein (eds.): *Distributions spatiales et temporelles*, 139-152, Benjamins: Amsterdam/Philadelphia, 1988.

[MARTIN 88b]
Martin, W., R. Heymans & F. Platteau: "DILEMMA, an automatic lemmatizer" in W. Martin (ed.): *COLINGUA 1*, UIA: Antwerpen, 5-62, 1988.

[MARTIN 89]
Martin, W. & G. Tops (eds.): *Van Dale Groot Woordenboek Engels-Nederlands*, Van Dale Lexicografie: Utrecht/Antwerpen, 1989, 2nd edition.

[MCNAUGHT 91]
McNaught J., B. Nkwenti-Azeh, W. Martin & E. ten Pas: *Feasibility of standards for terminological description of lexical items*, report written for the EC in the EUROTRA-7 framework, UMIST: Manchester Vrije Universiteit: Amsterdam, 1991.

[PALMER 90]
Palmer, M. S.: *Semantic processing for finite domains*, CUP: Cambridge, 1990.