Deriving *de/het* gender classification for Dutch nouns for rule-based MT generation tasks

Bogdan Babych
Centre for Translation Studies
University of Leeds
b.babych@leeds.ac.uk

Jonathan Geiger
Lingenio GmbH

geiger@cl.uni-heidelberg.de

Mireia Ginestí Rosell
Centre for Translation Studies
University of Leeds
mireia.ginesti@gmail.com

Kurt Eberle
Lingenio GmbH

k.eberle@lingenio.de

Abstract

Linguistic resources available in the public domain, such as lemmatisers, part-of-speech taggers and parsers can be used for the development of MT systems: as separate processing modules or as annotation tools for the training corpus. For SMT this annotation is used for training factored models, and for the rule-based systems linguistically annotated corpus is the basis for creating analysis, generation and transfer dictionaries from corpora. However, the annotation in many cases is insufficient for rule-based MT, especially for the generation tasks. In this paper we analyze a specific case when the part-of-speech tagger does not provide information about de/het gender of Dutch nouns that is needed for our rule-based MT systems translating into Dutch. We show that this information can be derived from large annotated monolingual corpora using a set of context-checking rules on the basis of co-occurrence of nouns and determiners in certain morphosyntactic configurations. As not all contexts are sufficient for disambiguation, we evaluate the coverage and the accuracy of our method for different frequency thresholds

in the news corpora. Further we discuss possible generalization of our method, and using it to automatically derive other types of linguistic information needed for rule-based MT: syntactic subcategorization frames, feature agreement rules and contextually appropriate collocates.

1 Introduction

This paper evaluates a methodology for deriving gender classification of nouns based on their contextual features and light-weight linguistic annotation of a corpus. We approach the problem as reconstructing an enriched set of linguistic features for RBMT generation lexicon from combining implicit information available in corpora with a set of general linguistic principles implemented as a small set of simple hand-crafted contextual rules.

These rules are specified as configurations of part-of-speech codes and operate over configurations of part-of-speech codes designed to capture certain disambiguating linguistic constructions. Theoretically, the rules can be made highly-accurate if the list of disambiguating constructions is exhaustive, but there is a well-known trade-off between Precision, Recall and the development effort for hand-crafted sets of rules. Additional factors to be taken into account are the quality and size of the annotated corpus. In our experiment we take a practical approach, using a minimal set of contextual rules that cover most typical constructions.

We evaluate Precision and coverage for this set of rules for different frequency thresholds of nouns in the corpus. The results indicate the potential of the proposed methodology for a larger set of similar tasks, where we intend to enrich linguistic resources for rule-based MT tasks using implicit linguistic information, which can be discovered in annotated corpora.

The paper is organised as follows: Section 2 discusses linguistic aspects of the gender disambiguation task for Dutch nouns; Section 3 describes the set-up of our experiment on automatically deriving the lexicon for Dutch nouns enriched with gender information; Section 4 presents evaluation results for Precision and coverage for different frequency thresholds; Section 4 gives interpretation of the results; Section 6 discusses the development context, generalisation of our methodology for rule-based MT and some ideas for future work.

2 Linguistic aspects of gender disambiguation task for Dutch nouns

Predicting gender of Dutch nouns from their context is a simple and clearly defined contextual disambiguation task, and we can evaluate three aspects of the performance of our method: (a) what coverage and accuracy can be achieved on this task compared to the gold standard; (b) how do the coverage and accuracy change in different frequency thresholds; (c) what is the proportion of contexts which can be used for disambiguation in different frequency thresholds (since some contexts will not disambiguate the features of interest).

Nouns in Dutch belong to one of the two gender classes which determine the choice of the definite articles (used with singular nouns) and other determiners: neuter nouns take determiners *het, dat, dit, ons*, and nouns with the common gender, which historically is the merged masculine and feminine, take *de, die, deze, onze*. Nouns can only be disambiguated when used as singular and take a definite determiner, so not all contexts in corpus which contain nouns can be useful for disambiguation.

The information about *het/de* classification for nouns is a non-interpretable (in terms of the generative grammar) system-internal morphological feature: it characterises only combinatorial properties of nouns, but does not directly influence their syntactic functions in a structure of a sentence or their semantic interpretation (unlike the part-of-speech/Noun category, morphological case and number). Therefore, this feature is much more useful for text generation than for analysis, and belongs to the family of other similar system-internal features, like inflection classes, subcategorisation frames, lexical functions (collocational restrictions), etc. Interestingly, this feature normally operates in the local context of several words within a limited number of possible part-of-speech sequences.

For machine translation task this information needs to be supplied by the target language generation rules, or by the target language model, since it is normally not present in the source text, and cannot be derived from application of transfer rules or the translation model.

There are several wide-coverage part-of-speech taggers and lemmatisers for Dutch in the public domain (open source and/or freely available), such as Dutch parameter files for the TreeTagger (Schmid, 1994), TiMBL / Frog tagger / lemmatiser / dependency parser (Van den Bosch et al., 2007), Alpino system (Bouma et al., 2001). Some of them provide only plain high-level annotation of part-of-speech codes, without gender information for nouns. However, some do generate enriched part-of-speech codes for nouns specifying their gender. Because of this we can benchmark our methodology using this enriched information as gold-standard and calculate Precision in addition to coverage.

3 Set-up of the experiment

In our experiment TiMBL / Frog was used to automatically annotate a 60-million-word section of the balanced Dutch SoNaR corpus (Oostdijk et al., 2008).

TiMBL/Frog provides gold-standard dictionary-based information about these classes for identified lemmas. For the prediction task we ignored the gold-standard gender class information, and used only the generic part-of-speech information and the number category for nouns. In the evaluation stage, we compared these automatically predicted gender classes with the gold-standard classes.

Prediction of the *de/het* classes was performed by a set of regular expressions, which cover most typical contexts, where these determiners are distinguished. If both types of determiners were found in different contexts for the same noun, then the class that has the majority of contexts was assigned. Regular expressions covered simple contexts, e.g.: *Det (Adj)? Noun*:

Frq threshold	Gold standard	Predicted	Wrong	%:100	Correct	%:100	Missed	%:100	Contexts %:100
None	157066	74505	2417	*0.032*	72088	*0.968*	84978	0.541	*0.752*
Frq>1	70006	45710	1604	*0.035*	44106	*0.965*	25900	0.37	*0.573*
Frq>2	48002	35766	1229	*0.034*	34537	*0.966*	13465	0.281	*0.518*
Frq>3	38084	30245	1012	*0.033*	29233	*0.967*	8851	0.232	*0.491*
Frq>4	32051	26515	858	*0.032*	25657	*0.968*	6394	0.199	*0.475*
Frq>5	28025	23818	744	*0.031*	23074	*0.969*	4951	0.177	*0.465*
Frq>6	25026	21735	661	*0.03*	21074	*0.97*	3952	0.158	*0.456*
Frq>7	22789	20053	597	*0.03*	19456	*0.97*	3333	0.146	*0.450*
Frq>8	21002	18701	543	*0.029*	18158	*0.971*	2844	0.135	*0.444*
Frq>9	19546	17553	498	*0.028*	17055	*0.972*	2491	0.127	*0.440*
...									
Frq>=20	12244	11436	279	*0.024*	11157	*0.976*	1087	0.089	*0.421*
Frq>=50	6795	6482	123	*0.019*	6359	*0.981*	436	0.064	*0.410*
Frq>=100	4297	4116	69	*0.017*	4047	*0.983*	250	0.058	*0.401*

Table 1. Evaluation of the task of predicting Dutch determiner classes: Number of tokens and proportions in each frequency threshold

(1) *de nieuwe geschiedschrijving*
the.Gend:COM new history.Gend:COM

-- but not more complex ambiguous contexts, e.g., sequences of nominal compounds:

(2) *waar is de apparaat-code van mijn ka-mera?*
Where is the~Gend:COM device~Gend:NEUT – code~Gend:COM of my camera?

or cases where *het* is not a determiner, but is mistagged as such: we assumed that such contexts are less frequent and error rate will be limited, so we can save the development effort for our hand-crafted rule set relying on the signal being stronger than noise introduced by such complex cases.

The results reported in this paper were generated using the following two multilevel regular expressions (expressions which operate on the levels of lemmas and parts-of-speech:

```
de/det__art     /(adj|conjcoord)*
(.*)/nounsg

het/det__art    /(adj|conjcoord)*
(.*)/nounsg
```

These regular expressions describe configurations that allow several optional adjectives or coordinative conjunctions between the definite determiner and a singular noun. The noun is captured if the configuration matches the piece of text and classified according to the type of the determiner.

4 Evaluation results

The results are presented in Table 1 and Charts 1 and 2, which visualise some of the data from Table 1.

Rows in Table 1 represent different frequency cut-off points, e.g: *None* = no frequency cut-off, *Frq>1* = noun types with frequency greater than one, etc. Columns represent:

- **Gold standard**: the number of noun *types* identified in the gold-standard above the specified frequency
- **Predicted**: the number of noun *types* for which prediction of the gender on the basis of the context in the corpus was made (for the rest prediction was not possible since no disambiguating contexts were found for those noun types)
- **Wrong, %/100**: the number and the proportion of wrongly predicted noun types (of the total number of **Predicted** types)

- **Correct, %/100**: the number and the proportion of correctly predicted noun types (of the total number of **Predicted** types)
- **Missed, %/100**: the number and the proportion of noun types where prediction of gender was not possible (of the total number of nouns in the **Gold standard**).
- **Contexts, %/100**: the proportion of contexts for noun *tokens*, which were useful for disambiguation

For instance, the first row shows the figures when no frequency cut-off is applied, e.g.: there were 157066 types labeled as Nouns in our section of SoNaR corpus, of which 74505 Nouns were found in a specific context with a definite determiner that allowed to disambiguate gender. Out of these, 2417 types (3.2%) were disambiguated wrongly for different reasons, 72088 types (96.8%) were disambiguated correctly. However, there still remain 84978 noun types (or 54.1% of the total number of 157066 in the gold standard), which were not disambiguated. In total, in the corpus 75.2% of contexts were useful for de/het disambiguation (contained a definite determiner in the immediate left context, or in a one-word-apart position, being separated by an adjective).

The second row in Table 1 presents the subset of 70006 noun types out of the results presented in the first row for 157066 noun types, i.e., the results only for nouns with frequency more than one; the third row – for noun types with frequencies more than two, etc. The intuition is that prediction for more frequent nouns should be more accurate since more *token* contexts become available for disambiguation of a specific noun *type*.
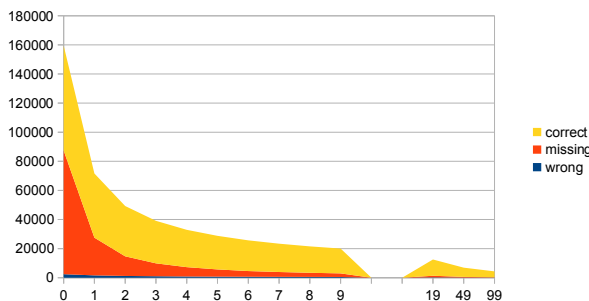

Chart 1. Distribution of correctly predicted, missed and wrongly predicted nouns

Chart 1 visualizes *correct, missing* and *wrong* proportion of noun types in the total count of these types for different frequency cut-off points. On the vertical axis there is a number of noun types, on the horizontal axis – *not greater than* frequencies.

It can be seen from the chart that the proportion of non-disambiguated noun types declines with increasing frequency threshold.
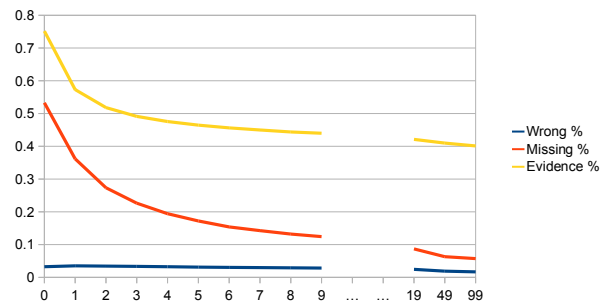

Chart 2. Proportion of context useful for disambiguation (evidence), not predicted (missing) and wrongly predicted (wrong) de/het classes for nouns.

Chart 2 examines the relation between frequency cut-off points and Evidence (top yellow/light line) – the proportions of contexts available for disambiguation; Missing (middle red/medium line) – the proportion of nouns where de/het disambiguation was not possible and Wrong (bottom blue/dark line) – the error rate.

5 Interpretation of the results

The following conclusions can be derived from the evaluation data:

1. The Precision even for simple contextual disambiguation rules is surprisingly high: 96.8% for nouns where the prediction was possible. This indicates that simple disambiguation patterns are sufficiently frequent to outweigh more complex patterns which were not covered by the rule and may have lead to errors.
2. For the whole data set (without frequency cut-off) the Recall is much lower: automatic prediction procedure missed 54% of noun tokens that were found in the corpus and a contained

gold-standard gender class, since no disambiguation context was found for these nouns in corpus. However, since more frequent nouns have more chances of occurring in a disambiguation context, mostly low frequent nouns are missed: if we exclude nouns which occurred only once, the procedure misses 37% of nouns; in frequency threshold Frq> 2 it misses even less – 28%, etc.

3. Error rate (proportion of wrongly disambiguated nouns) is relatively stable (3.2% on the whole data set), and does not depend too much on the frequency of nouns: it declines very slowly when the frequency increases (much slower that the coverage of the certain threshold).

4. The proportion of contexts which are useful for disambiguation declines slowly with the increase in frequency threshold, but stabilises around 40% for highly frequent nouns. Interestingly, when the proportion of such contexts goes down, the error rate stays the same.

In general, the results indicate that for practical purposes of rule-based MT development – a sufficiently large list of gender-disambiguated Dutch nouns (around 75000) can be successfully collected from a medium-size corpus (60MW) with very high Precision (96.8%). The method will provide gender disambiguation information for around 46% of all nouns found in the corpus; and for higher frequency threshold the percentage of gender-disambiguated nouns goes up rapidly, flattening at around 90% for Frq>10. This performance reaches the quality standards for creating wide-coverage generation dictionaries for rule-based MT.

6 Development context and generalization of the methodology

The task of predicting gender classes for nouns gives indication how other types of similar morphosyntactic resources and representations can be developed and enhanced.

Our methodology is part of a larger development infrastructure for creating a corpus-based development environment for industry-standard rule-based MT systems enhanced with statistical tools and data. The infrastructure uses large monolingual corpora annotated by openly available part-of-speech taggers and lemmatisers, and semi-automatically derives a set of morphological and syntactic patterns for the lexical items

found there. These patterns represent advanced linguistic features for the lexicon, such as classification by inflectional morphological paradigms, derivational classes (e.g., gender for nouns), lexical valencies (subcategorisation and case frames), attachment preferences and lexical collocates.

For individual lexical items these patterns do not need to be fully specified from the training corpus: missing forms are reconstructed on the basis of evidence from other lexemes that fit the same pattern, so the system recognises and generates correct output also for unseen forms.

In the context of our hybrid MT development infrastructure this approach particularly targets creation of linguistically-rich resources that generate correct target language forms and phrases. The generation aspect is usually not covered by the annotation tools available in the public domain, so parsers, part-of-speech taggers and lemmatisers usually work only in the direction of analysis, and do not deal with generation).

In a more general context the described infrastructure develops lexical and morphosyntactic resources in a systematic way, so they can be used in a wider range of applications and tasks. It also attempts to bridge the gap between rule-based and statistical techniques in MT by creating rich and highly accurate linguistic representations using corpus-based statistical techniques and integrating them within processing models for hybrid MT architecture.

The central principle of the proposed infrastructure is that advanced morphosyntactic features and representations are derived from corpora annotated with light-weight linguistic features.

The interpretation of this principle is that the tools like part-of-speech taggers and lemmatisers implement a unidirectional *functional perspective* on the morphosyntactic system, which only partially covers the network of linguistic relations involved in the analysis and generation aspects of the language. Rule-based MT application instead need to rely on the alternative *relational perspective* of morphosyntactic representations. Our infrastructure aims at reconstructing this perspective by combining large corpora and unidirectional annotation tools. It derives a range of generation-oriented morphosyntactic features and representations using local context and standard analysis-oriented annotation features in corpora.

The main motivation is that from the point of view of rule-based MT there is a certain imbalance between resources for analysis and annota-
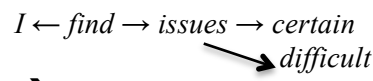
tion of texts on the one hand, and resources for language generation on the other hand. Text annotation resources, such as part-of-speech taggers, lemmatisers, parsers, chunkers – have a longer history of research and development, e.g., (Greene and Rubin, 1971), have created common standards and are more widely available in the public domain, e.g., (Schmid, 1994; Brants, 2000). In their existing form they can be applied to new languages and are more widely used in practical applications. On the other hand, generation-oriented tools are much less accessible, often propitiatory, and lack common standards and shared frameworks for integration of new languages. The predominant unidirectional text-annotation focus might be explained by a historic reason that text annotation was seen as an interesting computational problem with a clearly defined evaluation procedure, which was much harder to develop for the generation tasks.

The idea behind the infrastructure is that if at least some unidirectional annotation tools are available for a certain language, the relational morphosyntactic resources can be automatically developed from large annotated corpora. This will include automatic acquisition of inflectional paradigms for lexical items, attachment preference detection, automatic acquisition of lexical functions. Our infrastructure aims at developing standards and building openly available resources for a number of languages, including under-resourced languages, such as Portuguese, Russian and Ukrainian, in order to carry out the following morphosyntacitc tasks:

1. word form generation: for a given lemma, part-of-speech and inflectional feature values to generate the correct word form, e.g.: *drive~V + Person(3ʳᵈ); Number(singular) → drives*
2. generation of paradigms: for a given lemma and part-of-speech to generate a set of all word forms and their inflectional feature values, e.g., *drive~V → drive~VV; drives~VVZ; driving~VVG; drove~VVD; driven~VVN*
3. feature agreement generation: for a given sequence of lemmas with their part-of-speech codes to generate a correct sequence of inflected word forms, where inflectional features, e.g., in a language with adjectives and nouns marked for gender to generate a correct gender agreement between the two: in Spanish, e.g., *nuestro~A.Gender(_).Number(_)*

En:'our' + profesora~ N.Gender(fem).Number(plur) En:'professors(female)' → nuestras profesoras

4. lexical feature generation: to select correct lemmas for lexically underspecified structures, e.g., in a language with the gender feature marked on determiners and nouns to select the correct determiner to go with a given noun: Dutch: *[Determiner.Def(definite)] + beroep~N.Number(singular) → het beroep*
5. subcategorisation frame generation: to generate the correct prepositional phrase and/or morphological case features for a given verb and a noun (or a noun phrase), e.g.: *dispence~V + N → dispense with + N; dispose~V + N → dispose of +N*
6. collocate / lexical function generation (in terms of Mel'čuk, 1998): to select the correct lemma or ranked set of lemmas for a given word and semantic features of its context, e.g., '[not-real/true] + [Noun]': *mock trial; false assumption; counterfeit goods; fake name*
7. word order generation: to generate correct linear sequence of words for a given dependency structure, e.g.:

I ← find → issues → certain
↘difficult
➔
I find certain issues difficult

The first two functions are performed on internal features of a word, while the other five require contextual input in addition. The described functionality has applications for rule-based MT and Natural Language Generation, which could both benefit from shared standards and the infrastructure of relation-oriented linguistic resources.

Acknowledgement

References

Bouma, G., van Noord, G., and R. Malouf. (2001). Alpino: wide coverage computational analysis of Dutch. In W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel, editors, Computational Linguistics in the Netherlands 2000, pages 45--59. Rodolpi, Amsterdam.

Brants, T. (2000), TnT - A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA.

Greene, B. B. & Rubin, G. M. (1971), Automatic Grammatical Tagging of English. Department of Linguistics, Brown University, Providence, Rhode Island

Mel'čuk, I. A. (1998). Collocations and Lexical Functions. In Anthony P. Cowie (ed.) Phraseology. Theory, analysis, and applications, 23–53. Oxford: Clarendon.

Oostdijk, N., M. Reynaert, P. Monachesi, G. van Noord, R. Ordelman, I. Schuurman, V. Vandeghinste. From D-Coi to SoNaR: A reference corpus for Dutch. In: LREC 2008.

Schmid, H. (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Van den Bosch, A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting, Leuven, Belgium, pp. 99-114.