11-712: NLP Lab Report

Lingpeng Kong

April 26, 2013

Abstract

This is an incomplete draft of the development report for YAGMA (Yet Another German Morphological Analyser). The main content of this report includes basic information about German, past work on morphology of German, a survey of German morphology, design plans and experiment results for YAGMA.

[brief introduction –NAS]

1 Basic Information about German

German (Deutsch) is spoken as a first language by approximately 100 million people in the Federal Republic of Germany, the German Democratic Republic, Austria, Switzerland, and elsewhere, and as a second language by many others in Central and Eastern Europe. (Fox, 1990)

Most German vocabulary is derived from the Germanic branch of the Indo-European language family (European Commission, 2004). A number of words are derived from Latin and Greek, and fewer from French and English. German is written using the Latin alphabet. In addition to the 26 standard letters, German has three vowels with umlauts (Ä/ä, Ö/ö, and Ü/ü) and the letter ß.

In terms of morphological system, German has a particularly complex word structure, rather more complex than some other languages, including English. (Fox, 1990) It has a relatively speaking complex case and gender system, and very complicated rules in conjugation, declension and word formation.

2 Past Work on the Morphology of German

A lot of work has been done in the study of German morphology. Traditional linguistics provides a descriptive summary of the morphological phenomena in German (Boase-Beier, 2003) (Fox, 1990), which gives us insights of what is morphology in German, and how methods like "Umlaut" been used in word formation.

In the computational linguistics side, a handful of papers have been published (Lezius, 1996) (Lezius, 1998) (Lezius, 2000) (Schmid, 2004). These efforts fall into mainly different categories. The first one is to use a traditional stemming-then-generation approach, the basic idea of which is to cut all the possible affix and umlaut variations, then use the possible generation method to recover all the possible forms of the stem, if one (or more) of them match(es) the word given, we know its morphological components. The second approach is based on Finite State Transducer (FST), like much of the modern morphological analysers. These tools differs from each other mainly due to their

coverage of the morphological phenomena (some only consider inflections while others only consider derivations). SMOR (Schmid, 2004) is a very comprehensive morphological analyser which consider both inflections and derivations. It also make rules to deal with simple compounding phenomena in German and achieve a good result.

Despite the difficulties in German, it seems that it can still be solved in FST framework. Not much work (if any) tries to use upgrade to the context free grammar.

3 Available Resources

There are many available resources that can contribute to German morphological analysis tasks.

- Negra[1] and TIGER[2] are most widely used two Treebank for German. TIGER has morphological labels for the words since version 2.1.
- The CoNLL-2003 shared task data files[3] contain some labelled German text, where it give the words, lemma and its POS tags.
- Morphisto[4] is a morphological analyzer and generator for German wordforms. The basis of Morphisto is the open-source SMOR morphology for the German language developed by the University of Stuttgart (GPL v2) for which a free lexicon is provided under the Creative Commons 3.0 BY-SA Non-Commercial license.
- Vollformenlexikon[5] provides a German full form dictionary with about 90,000 base forms and 431,000 full forms (inflected forms). Each full form comes with full morphological information. The data is originally extract from Morphy (Lezius, 1996).
- LDC also provides several useful German corpus. (LDC95T11[6] is used for test for Schmid (2004))
- SFST - Stuttgart Finite State Transducer Tools [7]

In terms of the choice of our corpus A, B, C, we decided to use the TIGER Corpus since its labels are relatively high quality. We may also consider to mix some of the full forms in Vollformenlexikon because the TIGER labels are unique (words' morphological roles in a sentence are unique but only given full forms, we should have multiple possibilities for each word) while Vollformenlexikon labels all the possible morphological analysis of single full form word.

4 Survey of Phenomena in German

To make life easier, instead of talking countless exceptions and irregular rules, we will first introduce the most productive rules in German. Of course, this set of rules fall into the inflection category. These rules is the mainly focus of our work. After that, we will briefly discuss other morphological phenomena in German.

[1] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html
[2] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
[3] http://www.cnts.ua.ac.be/conll2003/ner/
[4] http://code.google.com/p/morphisto/
[5] http://www.danielnaber.de/morphologie/index_en.html
[6] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC95T11
[7] http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html

4.1 Case and Gender

German has a relatively complex case and gender system. It has four cases, the NOMINATIVE, ACCUSATIVE, GENITIVE, and DATIVE. It has three genders: MASCULINE, FEMININE, and NEUTER (in the plural, no distinction of gender is made).

4.2 Productive Morphological Phenomena (Inflection)

4.2.1 Nouns

Nouns can add some affix to generate their plural forms[8].
N [+SUFFIX (e)n, nen] → Plural
N [+SUFFIX s] → Plural
N [+SUFFIX e] ([+UMLAUT]) → Plural

Nouns can also have variations in cases.
Strong:
SIN-N [+SUFFIX (e)s] → Genitive
PLU-N [+SUFFIX (e)n] → Dative
Others remain unchanged.
Weak:
N [+SUFFIX (e)n] → All cases except Nominative

4.2.2 Verbs

Verb stems can add tense affix to generate infinite, past, past participle form.
Weak (Main Part of the German):
$[stem]_v$ + [+SUFFIX (e)te] → Past
[+PREFIX ge] + $[stem]_v$ + [+SUFFIX (e)t] → Past Participle

4.2.3 Adjectives

Adjectives can be various in cases, genders, and numbers[9].
Weak:
$[stem]_{adj}$ + [+SUFFIX e] → SIN MAS/NEU/FEM NOMINATIVE
$[stem]_{adj}$ + [+SUFFIX e] → SIN NEU ACCUSATIVE
$[stem]_{adj}$ + [+SUFFIX en] → Others

In German, adjectives also have their comparative and superlative forms. This is quite like in English.
$[stem]_{adj}$ + [+SUFFIX er] → Comparative
$[stem]_{adj}$ + [+SUFFIX (e)st] → Superlative

[8] A full discuss of these rules, like the choice of the rule based on different conditions will not be listed here to save space. Do note these rules exist, for example, when the noun ending with el, we should use the first rule rather than the second and third

[9] More precisely, when used with nouns in different cases, genders, and numbers, the adjectives have different forms.

4.3 Less Productive Morphological Phenomena

German has many exceptional and complicated rules (otherwise it will not considered to be thus difficult), but we won't worry too much about the irregular morphological changes, after all, they are finite, which means in worst cases, we can deal them with a table list all of them.

4.3.1 Derivations

Derivational, or lexical, morphology differs from inflectional morphology in being concerned with the formation of different lexemes from the same root, rather than with different grammatical forms of the same lexeme. Derivation is also very important in German, but they are less productive. Same thing happens in English. Some main phenomena related to derivation is summarized here[10].
[PREFIX ent/er/...] + $[stem]_v \rightarrow$ Verb
$[stem]_v$ + [SUFFIX eln/ern/...] $\rightarrow$ Verb
$[stem]_v$ + [SUFFIX er/...] $\rightarrow$ Noun
$[stem]_v$ + [SUFFIX bar] $\rightarrow$ Adjective
$[stem]_n$ + [SUFFIX ig] $\rightarrow$ Adjective
$[stem]_n$ + [SUFFIX isch] $\rightarrow$ Adverb

Conversion (Zero-derivation) also happens a lot in German, which always forms a word of a different category from the base word. But it is not really productive (Boase-Beier, 2003). Therefore, it seems that remember all of them in the finite state transducer is a reasonable solution for this.

4.3.2 Compounding

Compounding is one of the main difference between German and English. In German, compounding is every where. VV structure is quite rare, but we can still see many forms like NN NA AA AN and PN. This compounding may not only have the direct concatenate of two words, something may also be added in this process, like the following example:
$[stem]_n$ + s + $[stem]_n \rightarrow$ Noun
Fortunately, the good thing here, is that, after compounding, same rule set can be applied to it, the compounding noun has no particular features, it just performs as a common noun. Cases are also true for verbs and others.

4.4 Discussions and Notes

Despite the common affix things, German has some phenomena which worth more attentions. First is that, some morphemes are discontinuous, like the ge - en and ge - t. Second it that, a morpheme can be realized by suffix and umlaut together. These facts contribute to the complexity of the finite state transducer for German morphology analysis.

Other phenomena also do appear in German, like Suppletion, Ablaut, Chipping and Reduplication. We will not go through each of them here.

[10] Again, we do not tend to list all. We want to provide a general intuition about what happens in German.

5 Initial Design

5.1 Empirical Word Cluster

Rather than focusing on the morphological rules and way too many exceptions in German, our approach, like the in Schmid (2004)

5.2 Suffix Phenomena Solution

5.3 Deletion Phenomena Solution

For the deletion problem, unlike in English (e-deletion), we do not have a clear rule to do the deletion and the characters deleted are different from each other and often more than one character. Therefore, we use the different deletion rules to solve the problem, for example, if we want to delete the *a* in the end of the word, we implement an "Adel" (A-Deletion rule) to deal with that. Basically, if we first add a "-a" string before the morphological boundary (in the lexicon file), and then we use the rule to delete all the "a-a" before the morphological boundary (in the rule file).

```
define Delus {us} -> 0 || ?* _ "-us" "+um"* "^" ?* ;
```

5.4 Prefix Phenomena Solution

The prefix phenomena is solved using a trick FOMA regular expression which stands for a insertion before the longest possible match.

```
define Addpref   .?* "^" "+ge" ?*. @-> {ge} ...;
```

5.5 Umlaut Phenomena Solution

```
define Plusum1 a -> "+uma" || C* _ RC*  RuleSymbolPM* "+um" "^" ?* ;
```

First map the character into intermediate symbols like the "+uma", and then map it into the umlaut characters at last like "ä". By performing the trick, we avoiding the errors which happens if there are two different vowels in the same word.
Also, an open question is that, when there are two vowels in the same word, it is not clear for which one we should do the umlaut thing (It is neither always the first vowel nor the last vowel). Since we found that the vowel u seems to be more preferable in umlaut, we insert a tricky rule which replace the u first than a or o.

6 System Analysis on Corpus A

	Predicted	Corrected	Accuracy
Corpus A	573	560	0.978

	Predicted	Corrected	Accuracy
Corpus A	195	195	1.00

7 Lessons Learned and Revised Design

8 System Analysis on Corpus B

	Predicted	Corrected	Accuracy
Corpus B	650	615	0.946
	Predicted	Corrected	Accuracy
Corpus B	142	142	1.00
	Predicted	Corrected	Accuracy
Corpus A	657	631	0.960

9 Final Revisions

10 Future Work

References

European Commission, *Many tongues, one family. Languages in the European Union.* 2004

Anthony Fox, *The Structure of German*, Oxford University Press, 1990

Timothy Buck, *A Concise German Grammar*, Oxford University Press, 1999

Jean Boase-Beier, Ken Lodge, *The German Language - A Linguistic Introduction*, Blackwell Publishing, 2003

Yousheng Cai, Xiesong Jiang *Neue Deutsche Grammatik*, Shanghai Foreign Language Education Press, 1999

Wolfgang Lezius, Reinhard Rapp, Manfred Wettler, *A Freely Available Morphological Analyzer, Disambiguator, and Context Sensitive Lemmatizer for German*, in Proceedings of the COLING-ACL 1998 pp. 743-747

Wolfgang Lezius *Morphologiesystem Morphy*, in R. Hausser, editor, Linguistische Verifikation, Dokumentation zur ersten Morpholympics 1994 pp. 25-35, Niemeyer, 1996

Schmid, H., Fitschen, A., Heid, U. *SMOR: A German computational morphologycovering derivation, composition and inflection*, In: Proceedings of the 4th Inter-national Conference on Language Resources and Evaluation, Volume 4., Lisbon, Portugal (2004) 1263 - 1266

Wolfgang Lezius *Morphy - German Morphology, Part-of-Speech Tagging and Applications*, in Ulrich Heid; Stefan Evert; Egbert Lehmann and Christian Rohrer, editors, Proceedings of the 9th EURALEX International Congress pp. 619-623 Stuttgart, Germany