

Choosing the right lemma when analysing German nouns

Martin VOLK

1. Introduction

When processing large corpora, it is often necessary to lemmatise the wordforms. This is usually done by a morphological analyser which can, in any case, undo inflection but sometimes even derivation and compounding. The latter is especially useful for German which exhibits very productive compounding. But when using such a system we notice that lemmatisation is a frequent source of ambiguities. Some wordforms genuinely belong to two lemmas of the same part-of-speech such as *rasten* which can be a form of *rasen* ('to race') or *rasten* ('to rest'). Others belong to two lemmas of different word classes such as *meinen*, which can represent various forms of either the first person possessive pronoun ('my') or the verb 'to mean'. This latter ambiguity can easily be resolved by a part-of-speech tagger or a parser.

The former ambiguity is much harder to deal with. In the case of verbs a parser might be able to distinguish between the two lemmas if they subcategorise for different complements. It gets more difficult for nouns which often do not have clear subcategorization requirements. But our corpus studies show that nouns are a frequent source of ambiguous lemmas, in particular if different segmentations (compound and derivation segments) are taken into account. When analysing a newspaper corpus of the *Neue Zürcher Zeitung* we found that close to 10% of all noun types are assigned more than one lemma by our lemmatiser, the Gertwol system (HAAPALAINEN & MAJORIN 1994). The following examples show a number of ambiguous German noun forms whose lemma alternatives correspond to very different word meanings.

Abteilungen → (die) Abt~ei#lunge OR
(die) Ab|teil~ung

Ministern	→	(der) Mini stern	OR
		(der) Minister	
Flugzeuge	→	(der) Flug#zeug~e	OR
		(das) Flug zeug	
Verbrechen	→	(der) Verb#rechen	OR
		(das) Ver brech~en	

Some of these ambiguities may be resolved by using the gender information which can be inferred from the accompanying determiner. But for many others, this criterion cannot be employed since the variants show the same gender or the wordform occurs without a determiner or with an ambiguous determiner. We therefore investigated a method to use the segmentation information to decide on the correct lemma for German nouns. Our method relies on the segmentation information of the Gertwol system. Gertwol distinguishes four types of segmentation (quotes are from HAAPALAINEN / MAJORIN 1994, section 2.5.):

1. "Elements that can occur as independent words are separated with a strong boundary character (#). Verb stems occurring as first elements are still an exception to this rule. Examples: Berg#wiese, Schreib#maschine."
2. "Prepositions, prefixes and non-independent elements are separated with a weak boundary character (|)." Examples: vor|schule, geo|morpho|log~isch. The weak boundary is also used before non-independent second parts of the compounds. Examples: Mensch\en|recht~ler, zwei|j~ahr~ig. Half-suffixes ("a productive word whose meaning in compounds has changed from its meaning as an independent word") are also separated with a weak boundary. Examples: Laub|werk, but: Nach|schlag\e#werk.
3. "Linking elements may occur before the boundaries of compound words or suffixes. They are separated with a backslash (\). On the left-hand side of the linking element is the stem of the word." Examples: B~uch\er#bus, Fried~e\ns#freund.
4. Derivational suffixes are separated by a tilde character.

On the basis of these segmentations we have developed a disambiguation method that determines the correct lemma for about 90% of

ambiguous noun lemmas. It does not rely on tagging or parsing but only on the internal structure of the competing lemmas. The method is therefore well suited for shallow corpus investigations.

2. The disambiguation method

The disambiguation method is based on the observation that in most cases the noun lemma with the least internal complexity is the preferred lemma. A word's internal complexity depends on strong and weak composition boundaries as well as on derivation boundaries. If, for example, one lemma has a strong composition boundary and the other has a weak boundary, then the lemma with the weak boundary is preferred. The linking elements (interfixes) are phonologically motivated and do not influence the complexity.

The disambiguation method works in three steps.

1. Gertwol distinguishes regular nouns and derived nouns, where derived nouns are marked, if they are derived from adjectives (e.g. *das Gute*) or participles (*das Geschehene*, *der Sehende*).

Our first rule: If a noun has competing lemmas, where at least one lemma is marked by Gertwol as a regular noun, then discard all derivational noun lemmas. Example:

```
Hoffnungsträger →
    Hoffn~ung\s#träge  noun derived from adj.
    Hoffn~ung\s#träg~er  regular noun
```

2. Gertwol distinguishes strong and weak composition boundaries as well as derivational boundaries. These will be counted for every lemma according to the following scores:

- A strong composition boundary (#) gets 4 points.
- A weak composition boundary (|) gets 2 points.
- A derivation boundary (~) gets 1 point.

Our second rule: The lemma with the smallest overall point score is the best lemma. Examples:

```
Antragsteller →
    An|trag\s#teller = 6 points
    An|trag|stell~er = 5 points
```

Lohneinbussen →
 Lohn#ein#bus = 8 points
 Lohn#ein|buß~e = **7 points**
 Geldwäschereibestimmung →
 Geld#wäsch~e#reib~e#stimm~ung = 15 points
 Geld#wäsch~er#eib~e#stimm~ung = 15 points
 Geld#wäsch~er~ei#be|stimm~ung = **13 points**

A strong composition boundary counts more than the sum of a weak boundary and a derivation boundary, since these may lead to alternative lemmas. The above examples show that this difference helps to correctly discriminate between the alternative lemmas.

When manually checking the results on 400 ambiguous nouns, we found that our rules 1 and 2 lead to the correct lemma for around 85% of the ambiguously segmented noun lemmas. But we noticed that some of the remaining errors were due to some rarely used morphemes. Consider the following example where the rarely used word *Stag* (a hemp rope) is a possible compound segment.

Arbeitstag →
 Arbeit\s#tag = 4 points
 Arbeit#stag = 4 points

3. We therefore use lemma preferences to exclude the most exotic compound segments. We collected pairs of words that account for alternative lemmas where one word in the pair is clearly more unlikely to occur in the subject domain than the other.

Our third rule: If a ‘best’ lemma determined by our second rule ends with a dispreferred word (called ‘bad’ segment) and if an alternative lemma ends with a corresponding preferred word (called preferred segment) then accept the alternative as the best lemma.

With a list of 14 preference pairs our method improved to 90% correct lemmas for our newspaper corpus. Some examples from the preference list:

'Bad' segment	Preferred segment	Example
Buchs	Buch	Liederbuchs → Lieder#buch
Port	Sport	Motorsport → Motor#sport
Reis	Reis~e	Ferienreise → Ferien#reis~e
Samt	Amt	Arbeitsamt → Arbeit\s#amt
Stag	Tag	Arbeitstag → Arbeit\s#tag
Tand	Stand	Wohlstand → Wohl#stand
Tuba	Stube	Badestube → Bad\e#stube

Obviously the preference list must be adapted to the text type. For instance, analysing our newspaper corpus we found it advantageous to prefer *Reis~e* ('trip') over *Reis* ('rice'), whereas for a cookbook the opposite will be true.

3. Comparison to WordManager

One may object that the above rules solve a problem that only arises due to Gertwol's way of morphological analysis (dynamic undoing of compounding and derivation). A competing system like WordManager (DOMENIG & HSIUNG 1996) provides the correct lemmas for our examples (*Abteilungen*, *Ministern*, *Flugzeuge*, *Verbrechen*). But Word Manager achieves this perfect result at the cost of underanalysis. It is unable to analyse *ad-hoc* compounds like *Schweinelunge*, *Unfallzeuge*, or *Heurechen*.

Gertwol's approach is much more flexible and will thus achieve a higher lexical coverage of a natural German text that will always contain newly created compounds.

4. Extension to German verbs

In recent experiments we observed that our approach, first developed for nouns only, carries over to verbs with only little modifications. In a study on a *ComputerZeitung* corpus (one year's issues summing up to

about 1.5 million tokens) we found around 10,000 verb form types. Gertwol finds a unique verb lemma to about 8,700 of these verb forms. It cannot find any lemma for around 700 verb forms, most of which are English words, some containing spelling errors. This leaves around 600 cases with more than one lemma. This is a surprisingly high number, since verb compounding is by far less productive in German than noun compounding. Here are some examples:

abgehandelt	→	ab handel~n OR ab hand~eln
bedacht	→	be denk~en OR be dach~en
gegenübersieht	→	gegenüber seh~en OR gegen über seh~en
mitentwickelt	→	mit entwickel~n OR mit ent wickel~n

Competing verb lemmas are due to a number of causes. They can be based on different nouns (Handel vs. Hand), different verbs (denken vs. dachen), complex or concatenated prefixes (gegenüber vs. gegen|über), or lexicalised vs. concatenated prefix-verb combinations (entwickel~n vs. ent|wickel~n).

Again our disambiguation relies on the principle of least internal complexity and on lemma preferences. Lemma preferences are more important for verbs than for nouns. Some examples:

'Bad' segment	Preferred segment	Example
dach~en	denk~en	bedacht → be denk~en
dring~en	dräng~en	verdrängt → ver dräng~en
fäll~en	fall~en	wegfällt → weg fall~en
fahr~en	führ~en	zurückführt → zurück führ~en
gerat~en	rat~en	abgeraten → ab rat~en
kos~en	kost~en	auskosten → aus kost~en
miss~en	mess~en	beimißt → bei mess~en

Lemma preferences will be applied in case of competing verb lemmas. If, for example, a verb form can have both a lemma with the stem *fahr~en* and with the stem *führ~en* then the latter is used. Note that the stem of the preferred segment is often the conjunctive stem for the ‘bad’ segment, a form that is seldom used in German.

5. Conclusion

We have developed a method to find the correct noun lemma in cases of segmentation ambiguity. The method is based on local heuristics which use the Gertwol composition and derivation boundaries. The method has been implemented in Perl. The program can be used as a filter on the Gertwol output, thus reducing the ambiguity for further processing steps.

As a refinement we will use frequency information of the segments to improve the preference list. That is, we will check for all compound segments how often they occur in a given corpus, and build the preference list automatically according to the most frequent segments. In this way we will extend our method to global corpus information.

References

- DOMENIG, Marc and HSIUNG, A. (1996): Concepts and tools for lexical knowledge acquisition. In: *AI communications* 9(2), 79-82. (<http://www.wordmanager.com>).
- HAAPALAINEN, Mariikka and MAJORIN, Ari (1994): GERTWOL: Ein System zur automatischen Wortformerkennung deutscher Wörter. Lingsoft, Inc. (<http://www.lingsoft.fi/cgi-pub/gertwol>).