

Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values

Uwe Quasthoff

Universität Leipzig

D-04109 Leipzig, GERMANY

[quasthoff@informatik.uni-leipzig.de]

Abstract

The paper describes the algorithmic methods used in a German monolingual lexicon project dealing with a multimillion entry lexicon.

We describe the usability of different information which can be extracted from the lexicon: For German nouns and adjectives, candidates for their inflexion classes are automatically detected. Forms which do not fit in these classes are good error candidates. A n-gram model is used to find unusual combinations of letters which also indicate an error or foreign language entries. Regularity is used especially for compounds to get inflection information. In all algorithms, frequency information is used to select terms for correction.

Quality information is attached to all entries. Generation and use of this quality information gives an automatic control over both the data and the correctness of the algorithms.

The algorithms are designed to be language independent. Language specific data (as inflexion classes and n-grams) should be available or relatively easy to obtain.

Introduction

In the monolingual project "German Vocabulary" (Quasthoff, 1998) a lexicon of full form German words was automatically generated using machine readable corpora, mainly news papers and scientific journals. The project started in 1996. Reading about 250.000.000 words resulted in a word list containing approximately 3.700.000 entries. Between one and two percent of the entries contain errors, mainly due to spelling errors in the source text and misinterpretation of layout information.

Because of the large number of entries there is a strong demand for algorithmic error correction methods.

The lexicon in its current state is available via internet at <http://www.wortschatz.uni-leipzig.de>.

Lexicon acquisition

Most of the corpora to be analysed for lexicon build-up arrived on CD-ROM in various formats. After converting them to ANSI format we perform a sentence segmentation. Next, the words are segmented and looked up in the lexicon. If a certain word is found, its absolute frequency is updated. If it is not found, it is appended to the lexicon together with the sample sentence it was found in. There is a special treatment for the first word in each sentence to solve the problem of lower case / upper case conflicts.

Moreover, each lexical entry comprises the following information (if available):

1. If the word is in basic form, its inflection class is (or several possible classes are) stored.
2. If the word is an inflected form, the inflection rule is given. This also serves as a hyperlink to the corresponding basic form.

This information can clearly not be extracted from running text, but from the whole collection assuming it contains all inflected forms.

Using other sources like telephone directories or technical language word lists we also gather additional information like proper name tags or subject areas.

Moreover, using available machine readable dictionaries we have grammatical information (part of speech, inflection) for an initial set of basic forms.

Automatic detection of inflection classes

The collection of words described above can be used in two different ways for finding the inflection classes. The first method uses the information given by the fact that probably all inflected forms of a given basic form are contained in the lexicon. The second method simply assumes that similar words inflect in a similar kind. Of course, success depends heavily on the kind of similarity used.

Checking possible inflected forms

For every word in basic form, all possible inflected forms are generated. We do not use any extra knowledge but for a noun, for instance, simply add noun suffixes, perform German Umlaut transformation etc. Every generated (i.e. possibly existing) word is checked against the lexicon to decide upon its existence. From the set of existing inflected forms the inflection class is derived. Here we use a complete list of German inflection classes (similar to Maier-Meyer (1995)) and frequency information for the inflected forms. In addition to the suggested inflection class (which is mostly correct if the frequency exceeds a certain minimum) we find existing inflected forms of low frequency which do not fit into inflection class. These are good error candidates.

Example: For the German word *Anflug* (approach of an air plain) the database contains the following forms (given with frequency information): *Anflug* (404), *Anflug-e* (2), *Anflüg-e* (36); *Anflüg-en* (12); *Anflug-s* (7). Using this information it is concluded that *Anflug* inflects like *Baum* and the word *Anfluge* is possibly incorrect.

Inflection classes based on similarity

This algorithm is mainly designed for German compounds in the case the inflection class of the last component of the compound is known. The problem here is to detect the correct decomposition of a compound. Moreover, it is possible to use the same method if a word ends in a certain string which determines the inflection class.

Using morphological information. We use a morphological analyser which gives results as described in (Wothke, 1993). Compounds are segmented by =, suffixes by %.

Example: *Modehaus* (fashion store) is segmented as =*mod*%*e*=*haus* and therefore inflects like *Haus* (which is assumed to be known).

Using characteristic word ends. The following algorithm is weaker because we do not use morphological information, but it can be applied in a more general context. If all words ending in the same string (for instance, *-bung*) and for which the inflection class is known are in fact in the same inflection class, then this inflection class will be assigned to all words ending the same.

This algorithm heavily relies on sufficiently many initial data to start. If, for instance, it is not previously known that *Schwung* or *Dung* are masculine, it would conclude that all German words ending in *-ung* are feminine (which is the case for almost all words ending in *-ung*). Hence, this algorithm should be applied very carefully and the results should be checked if possible.

N-gram models

There are a lot of pairs, triples, or quadruples (2-, 3- or 4-grams, see (Salton, 1998)) of letters which are very unlikely to appear in a German word. They can be used to detect either orthography errors in German words or to identify non-German words.

Error detection and correction

For such words one can look for similar words (for instance, differing in just one letter) which do not contain the exceptional sequence. If this word is contained in the lexicon and has sufficient high frequency, it is a good candidate to be the correct version of the original word. This procedure should, of course, not be applied to proper names.

Examples: The trigram *wss* is not allowed in German words. Hence, *Wssenschaft* (frequency 1) turns out to mean *Wissenschaft* (6696). Similarly, there are no German words ending in *-nb*. Hence, using the sample sentence, the word *Vereinb* (1) turns out to be the abbreviation for *Vereinbarung* (93230). In the first example we allow automatic correction. In contrast, the second example requires human interaction.

Detecting foreign language words

A trigram method is used to detect foreign language words. Because foreign language detection is impossible at word level for many words, we use the sample sentences instead. If a sample sentence is found to be in

non-German, the words contained in this sentence are marked as probably foreign. In many cases we can determine the language of a sentence using a trigram model and mark the words as possibly English or Latin, for instance.

Of course, a certain word (for instance, *hat*) can be marked both as possibly English and as correct German. For marking words as correct German see below.

Pre-processing for manual control

There are lots of words which might contain an error but there is not enough certainty for an automatic replacement. In this case the words are marked as possibly incorrect. This information is stored for later decision either by human inspection or because of additional information.

In the case of a human inspection, the sample sentence usually gives the desired information.

Similarity to words of higher frequency

Many spelling errors cannot be detected using trigrams. But typically spelling errors have a frequency of at least a factor of 100 lower than the correct form. The problem here is to distinguish a spelling error from a correct, but rare word which is similar to a given correct word of higher frequency. Here one can use the traditional spell checker methods which look for typical typing errors first. Moreover, we can use semantic information. For instance, we usually have the information whether a word is a proper name which is rarely inflected. Omitting these words together with probably non-German words increases the performance significantly.

Incomplete data for high frequency terms

For an increasing number of high frequency words we want to assure maximum quality. Hence, high frequency terms with missing grammatical information will be offered for manual completion.

Using quality information

Quality information as described below is also used for selecting words for manual inspection. Candidates are high frequency words with negative quality information or words with different and inconsistent quality information.

Quality information in the lexicon

Storing quality information

Quality information is important for the further use of the corresponding data. Quality information is available as well for the word as for every information about a word (for instance, its inflection class). A quality entry contains a quality value and the reason of the entry, which can be either an algorithm or an expert source.

Entries with only negative quality information are not deleted. Otherwise, let us assume a misspelled word is deleted. In the case of a typical spelling error, this misspelling will appear again in analysed text and regenerate the old entry without the negative quality rating! But, of course, these entries are no longer visible to the standard user.

Positive quality information. Such entries are the result of a global or partial quality check. Typically a global positive quality information can only be given by human inspection or can be automatically implied using several partial quality checks. Again, the source of the quality information is important for a possible later refinement of the automatic implication.

The presence of grammatical data (inflection type, morphological decomposition) can be viewed as positive quality information.

Negative quality information. The various algorithms for finding possible errors are all sources for negative quality information. We have to distinguish between definite errors and much weaker signs for possible errors. There may be several indicators for a word to be incorrect which should imply a higher degree of falsity.

The absence of grammatical data (inflection type, morphological decomposition) can be viewed as negative quality information. This is the case especially for higher frequency words because they should possess these information.

Sources for Quality information

Quality information results from various processes and sources. Using both internal and external sources we get in many cases enough information for a reliable overall quality result.

Internal sources. The following sources are used within the lexicon project:

- Human inspection (i.e. editing the entry if necessary), afterwards the entry is marked as correct.
- Results of algorithms checking correctness of some kind. Correctness of the inflection class can be checked by comparing the given inflection class (which might be derived using morphology or similarity) with the existing inflected forms.

External Sources. External data and linguistic algorithms available from other projects are used to compare the lexical data. The following sources are used:

- High quality machine readable sources with earlier human inspection.
- Comparison with other word lists and use of external programs which accept 'correct' words according to different criteria.

Dealing with inconsistent quality information

In many cases the quality information as described above is not sufficient or even inconsistent. Here we are looking for intelligent procedures to decide the following questions:

- When can one quality information override another, contradicting information?
- Which combination of positive quality information assures correctness?
- Which combination of negative quality information assures incorrectness?
- How do we find rare errors possibly made by a special algorithm?

Example: The positive information that a special string like *dpa* is an abbreviation overrides the negative

information that the trigram *_dp* (*_* marks the beginning of a word) is not allowed in German.

Implementation aspects

The lexical database

Because of the large number of entries a relational database on a UNIX server is used. For looking up a special word we have the following WWW interface:



Figure 1: WWW interface for lexicon lookup

It is possible to use wildcards. In this case, one or more additional pages are generated which show all words fulfilling the search criterion. The lexicon lookup is available since march 1998. Single lexicon entries can be edited using a HTML form. For more complex lookups or complex changes one can use the more powerful SQL interface.

Sample entries

In the following we give two sample entries.

Entry for Weltanschauung. The following entry shows frequency, morphology, grammatical information and inflection class(es). In this example we have two possible inflection classes which differ in the existence of the plural form. These two inflection classes should be unified by allowing the plural form.

Wort: Weltanschauung

Häufigkeit: 392

Morphologie: =welt+an=schauung

Grammatikangaben:

Wortart: Substantiv

Geschlecht: weiblich

Flexionsklasse: fb, ff

Beispiel: Damals entwickelte sich daraus eine mechanistische Weltanschauung. (Quelle: Computerzeitung 1993)

Entry for *sang*. In the case of an irregular verb form the reference to the basic form is explicitly given.

Wort: sang

Häufigkeit: 1263

Grammatikangaben:

Wortart: Verb

Stammform: singen

Beispiel: Er sang mit kerniger Stimme,
aber betörend jüngerhaftem
Charme Wolframs Lied an den
Abendstern: "Die Todesahnung,
Dämmerung deckt die Lande".
(Quelle: FAZ 1994)

Bibliographical References

- Maier-Meyer, P. (1995). Lexikon und automatische Lemmatisierung; CIS-Report 95-84, Universität München.
- Quasthoff, U. (1998). Projekt Der Deutsche Wortschatz; in: Heyer, G., Wolff, Ch., *Linguistik und neue Medien*, Wiesbaden, Dt. Universitätsverlag (to appear).
- Salton G. (1989). Automatic Text Processing - The Transformation Analysis and Retrieval of Information by Computer, Addison Wesley.
- Wothke, K. (1993). Morphologically based automatic phonetic transcription, IBM Systems Journal 32 (pp. 485-511).