A comparison of sub-word indexing methods for information retrieval

Johannes Leveling
Centre for Next Generation Localisation (CNGL)
Dublin City University, Dublin 9, Ireland
johannes.leveling@computing.dcu.ie

Abstract

This paper compares different methods of sub-word indexing and their performance on the English and German domain-specific document collection of the Cross-language Evaluation Forum (CLEF). Four major methods to index sub-words are investigated and compared to indexing stems: 1) sequences of vowels and consonants, 2) a dictionary-based approach for decompounding, 3) overlapping character n-grams, and 4) Knuth's algorithm for hyphenation.

The performance and effects of sub-word extraction on search time and index size and time are reported for English and German retrieval experiments. The main results are: For English, indexing sub-words does not outperform the baseline using standard retrieval on stemmed word forms (–8% mean average precision (MAP), –11% geometric MAP (GMAP), +1% relevant and retrieved documents (rel_ret) for the best experiment). For German, with the exception of n-grams, all methods for indexing sub-words achieve a higher performance than the stemming baseline. The best performing sub-word indexing methods are to use consonant-vowel-consonant sequences and index them together with word stems (+17% MAP, +37% GMAP, +14% rel_ret compared to the baseline), or to index syllable-like sub-words obtained from the hyphenation algorithm together with stems (+9% MAP, +23% GMAP, +11% rel_ret).

1 Introduction

Splitting up words into sub-words is a technique which is frequently used to improve information retrieval (IR) performance. The main idea behind sub-word indexing is to break up long words into smaller indexing units. These indexing units can be found by methods such as decompounding words into lexical constituent words or splitting words into character n-grams of a fixed size. In some languages like German, compounds are written as a single word. Thus, if a German query or document contains a compound word like *"Kinderernährung"* (nutrition of children), the words *"Kind"* (child) and *"Ernährung"* (nutrition) will not match and result in low recall. Splitting the compound word and finding smaller indexing units will make a match more likely and yield a higher recall. For instance, a decompounding process may identify the constituent words *"Kinder"* (children) and *"Ernährung"*, which can be used in a query to achieve a higher IR performance. Linguistically oriented approaches aim at breaking up compound words into constituent words. Other approaches to generate sub-words do not build on the notion that sub-words must be valid words of the language (e.g. character n-grams).

For languages with a rich morphology (like Finnish, Dutch or German), a linguistically motivated decomposition of words has been widely recognised as a method to improve IR performance [Braschler and Ripplinger, 2003; Chen and Gey, 2004]. In languages such as English, compounds are typically written as separate words and their constituents can be easily identified.

However, creating resources such as dictionaries is expensive and time-consuming and dictionaries depend on language and domain. The most extreme knowledge-light approach at decompounding, overlapping character n-grams, has extreme requirements for index space due to combining grams for different values of n [McNamee, 2001; McNamee and Mayfield, 2007]. Decompounding methods should in the best case be efficient and effective, i.e. they should be inexpensive (i.e. not rely on external resources), largely independent of a particular domain; and adaptable to many languages. One aim of this paper is to help identify such an approach for decompounding words.

The contribution of this paper is the quantitative evaluation of four different sub-word indexing methods. The performance of the methods and their combination with stemming is compared for a compounding and a non-compounding language i.e., German and English. Sub-word indexing based on consonant-vowel-consonant sequences has primarily been used in speech retrieval and not in domain-specific information retrieval. Two of the variants of this approach (consonant-vowel sequences and vowel-consonant sequences) are novel. Knuth's algorithm for hyphenation has not been applied before to identify syllable-like sub-words as indexing units. Effects of sub-word indexing on the index size and on indexing time and search time are rarely discussed.

The rest of this paper is organised as follows: Section 2 introduces the sub-word identification techniques used in the IR experiments in this paper. Section 3 gives an overview over related work where approaches to decompounding have been employed. Section 4 describes the experimental setup for the experiments. Section 5 discusses the influence of sub-word indexing on retrieval performance, search time, and indexing time and space and provides a topic analysis. Section 6 concludes with a description of future work.

2 Identifying sub-words

The information retrieval experiments describes in this paper are conducted on German and English queries and documents to investigate the performance of sub-word identification for a compound-rich and a non-compounding language. Four different approaches to sub-word indexing are evaluated and compared to the baseline of indexing stems (stem):[1]

1. consonant-vowel sequences (CV) and derived methods, including vowel-consonant sequences (VC), consonant-vowel-consonant sequences (CVC), and vowel-consonant-vowel sequences (VCV);
2. a dictionary-based approach to identify constituent words of compound words (DICT);
3. syllable-like character sequences determined by Knuth's algorithm for hyphenation (HYPH); and
4. overlapping character n-grams (3-grams, 4-grams, and 5-grams).

Table 1 shows results of applying sub-word identification to the German word *"Informationssuche"* (information retrieval). The following subsections provide a more detailed description of these sub-word indexing techniques.

2.1 Dictionary-based decompounding

Dictionary-based decomposition of a word typically involves repeatedly determining whether prefix strings of a compound are valid words by looking them up in a dictionary. Many decompounding approaches used for German IR consider only the most frequent rule or rules of word formation. For example, the word *"Betriebskosten"* (operating costs) consists of two constituents, *"Betrieb"* and *"Kosten"*, connected by a so called Fugen-*s*. This connection represents one of the most frequent patterns in German compound word formation.

Dictionary-based decompounding is quite robust to some linguistic effects in the German language. For example, some compounds contain constituents in their plural form (e.g. *"Gänsefleisch"* (literally: geese meat)), which will be normalised to the same base as the words in singular form after stemming is applied (e.g. *"Gans"* (goose) and *"Fleisch"* (meat)). Some compounds should not be split into their constituents at all (e.g. *"Eisenbahn"* (railway) oder *"Lieblingsgetränk"* (favourite drink)), but these cases are rare and can be treated by using exception lists. Decompounding even allows for ambiguous results for the same compound. For example, *"Arbeitsamt"* (employment bureau), can be split into *"Arbeit"* (work), Fugen-*s*, and *"Amt"* (bureau) or into *"Arbeit"* and *"Samt"* (velvet). Ambiguities are typically resolved by a left-to-right, longest match preference.

However, dictionary-based decompounding requires language-specific dictionaries and additional processing time for successively looking up potential constituents in the dictionary to determine if they form valid words.

2.2 Consonant-vowel sequences

The Porter stemming algorithm [Porter, 1980] is a rule-based heuristic to normalise words to index terms by suffix removal. As a by-product, it computes the M-measure, a count roughly corresponding to the number of syllables in the word.[2] The M-measure is defined via the number of consonant-vowel-consonant sequences (short: CVC sequences) in a word. The set of vowels differs from language to language: In German, vowels are *"a"*, *"e"*, *"i"*, *"o"*, *"u"* (not counting letters with diacritical marks); in English, vowels also include *"y"* if preceded by a consonant. Other languages such as Arabic or Hebrew have no letters to represent vowels. The computation of the M-measure in the Porter stemmer can be easily adapted to generate sub-words, i.e. by adding a sub-word to a list each time M is increased. The M-measure can also be calculated for words in other languages by defining the corresponding set of vowels. The Snowball string processing language[3] provides stemmers for a range of different languages.

A CVC sequence is the longest match of a sequence of zero or more consonants (C), followed by zero or more vowels (V), followed by one or more consonants in a word. Three variants of these character sequences can be defined accordingly (VCV, CV, and VC sequences) and are investigated in this paper, too.

From an IR perspective, CVC sequences offer a cheap alternative to a complex morphologic analysis of words. As stemming has become a standard approach to normalise indexing terms, the modification of a stemmer to produce CVC sequences would require little additional cost.

2.3 Overlapping character n-grams

Words can be broken up into sequences of characters of a fixed size n to form character n-grams. If $n-$grams are allowed to start at every character position (instead of one n-gram for every n characters), the n-grams will partially overlap. Some variants of this method include adding an extra character as a special word boundary marker to n-grams from the beginning and end of a word. Following this approach and the character *"|"* as a boundary marker, the set of 4-grams for the noun *"Lichter"* includes the gram *"|lich"* from the beginning of the word and allows to distinguish it from the common adjectival ending *"lich|"*.

In another approach, the full text is regarded as a single string and not broken down into words before calculating n-grams. Whitespace characters are not discarded and become part of the character n-grams, which can span word boundaries.

2.4 Knuth's hyphenation algorithm

Knuth's hyphenation algorithm was developed by Knuth and Liang for dividing words at line breaks for the TeX/LaTeX typesetting tool [Liang, 1983; Knuth, 1984]. It is well documented and has been used in the document formatting system groff, in the PostScript language, and in the programming language Perl. At its core are sets of language-specific patterns. The patterns are employed to identify positions at which a line break can occur and a word can be divided. In this paper line break positions between two characters are interpreted as positions marking sub-word boundaries for sub-word identification,

3 Related Work

Decompounding is a successful method to improve retrieval performance in IR. There have been numerous re-

[1] Stemming can be viewed as a way to identify a single sub-word within a word by affix removal and is considered as a baseline for sub-word indexing.

[2] In a pre-test, the number of syllables was calculated correctly in about 93% using the M-measure on a test set of about 30,000 manually annotated words. Most errors resulted from foreign expressions and proper nouns.

[3] http://snowball.tartarus.org/

Table 1: Examples for splitting the German word *"Informationssuche"* into sub-words with different methods.

method	sub-words	# sub-words
stem	informationssuch	1
CV	i, nfo, rma, tio, nssu, che	6
VC	inf, orm, at, ionss, uch, e	6
CVC	inf, nform, rmat, tionss, nssuch	5
VCV	info, orma, atio, onssu, uche	5
DICT	information, suche	2
HYPH	in, for, ma, ti, ons, su, che	7
3-grams	inf, nfo, for, orm, rma, mat, ati, tio, ion, ons, nss, ssu, suc, uch, che	15
4-grams	info, nfor, form, orma, rmat, mati, atio, tion, ions, onss, nssu, ssuc, such, uche	14
5-grams	infor, nform, forma, ormat, rmati, matio, ation, tions, ionss, onssu, nssuc, ssuch, suche	13

trieval experiments using simple rule-based or dictionary based approaches to decompounding German words. Note: Most researchers report performance gain comparing sub-words originating from stems to a baseline with indexing unprocessed word forms. This results in better performance values (as effects of stemming are included in sub-words experiments), but make a comparison with other retrieval experiments more difficult.

Kamps et al. perform information retrieval experiments including decompounding to documents from the CLEF 2003 collection in nine languages. They report a 7.5% increase in MAP for an experiment on the German document collection including dictionary-based decompounding over baseline with stems and a 13.0% increase for 4-grams [Kamps *et al.*, 2003]. Results for decompounding English documents are not given.

Chen and Gey use dictionary-based decompounding to the CLEF 2001 and 2002 test collections [Chen and Gey, 2004]. Decompounding is based on computing the probability of the best splitting sequence based on the frequency of constituents [Chen, 2003]. For monolingual German retrieval experiments, they report a 12.7% increase in MAP and 4.6% in relevant retrieved documents for the 2001 data (13.8% and 13.1% for 2002 data, respectively) when indexing stemmed compounds together with their constituents compared to an experiment using only stems.

Daumke et al. apply MorphoSaurus as a text processing tool to documents [Daumke, 2007; Daumke *et al.*, 2007]. MorphoSaurus breaks down words into sub-words based on a dictionary with pseudo-morphological word elements. The sub-word segmentation of a word is determined automatically based on a manually created list of sub-words. For the English OSHUMED test collection, they achieve 5% increase in MAP compared to a stemming baseline; for German GIRT data, a decrease of 19.5% in MAP, and for German data from the image retrieval task ImageCLEF, an increase from 0.0343 to 0.0403 MAP (+17.5%).

Glavitsch and Schäuble extract CVC sequences as indexing features for retrieval of speech documents [Glavitsch and Schäuble, 1992; Schäuble and Glavitsch, 1994]. They select features based on document and collection frequency, and discrimination value. This indexing method performs slightly better than one using stopword removal and stemming. Similarly, Ng performs experiments on spoken documents for English, achieving 28% performance increase when combining sub-words indexing with error compensation routines [Ng, 2000]. CVC sequences are often used as indexing units for speech retrieval, even for non-European languages.

Braschler and Ripplinger give an overview about stemming and decompounding for German [Braschler and Ripplinger, 2003]. They perform IR experiments on data from CLEF for the ad-hoc retrieval track. They apply a variety of approaches for stemming and decompounding – including commercial solutions – and achieve a performance gain of up to 60.4% MAP and 30.3% for the number of relevant retrieved documents in comparison to indexing raw word forms (not stems).

McNamee performs retrieval experiments using overlapping character n-grams as indexing units [McNamee, 2001]. He reports performance results for indexing a combination of 2-grams, 3-grams, and 4-grams for English, Chinese, Japanese, and Korean. Results show that n-grams can achieve similar or superior performance in comparison to standard indexing techniques, even for non-compounding languages and for cross-lingual retrieval [McNamee and Mayfield, 2007].

To the best of the authors' knowledge, hyphenation algorithms or syllabification have not been applied to find sub-words for information retrieval on written documents before.

4 Experimental Setup and System Description

The retrieval experiments in this paper are based on data from the German Indexing and Retrieval Test database (GIRT) [Kluck, 2005] used in the domain-specific track at CLEF (Cross Language Retrieval Forum). The document collections in German and English consist of 151,319 documents from the GIRT4 database.[4] The topics include the 150 German and English topics from the domain-specific track at CLEF from 2003 to 2008 (25 topics each year), together with official relevance assessments.

A GIRT document contains metadata on publications from the social sciences, represented as a structured XML document. The metadata scheme defines 14 fields, including abstract, authors, classification terms, controlled terms, date of publication, and title. Figure 1 shows an excerpt from a sample document.

A GIRT topic resembles topics from other retrieval campaigns such as TREC. It contains a brief summary of the information need (topic title), a longer description (topic description), and a part with information on how documents are to be assessed for relevance (topic narrative). Retrieval

[4] In 2006, 20,000 abstracts from Cambridge Scientific Abstracts were added to the English GIRT document collection. As there are no relevance assessments available for topics from before 2006, these documents were discarded for the experiments.

queries are typically generated from the title (T) and description (D) fields of topics. Figure 2 shows a sample topic.

For each GIRT topic, relevant documents have been assessed by pooling submissions from systems participating in the domain-specific track at CLEF, resulting in a total of more than 80,000 relevance assessments for German documents (68,000 for English documents, respectively), including 16,200 German relevant documents for 150 topics (14,162 for English). The experimental results in this paper are based on the complete set of German and English topics and their corresponding relevance assessments.

The experiments were conducted with the following system setup. Lucene[5] was employed to preprocess the topics and documents and to index and search the document collection. The document structure was flattened into a single index by collecting the abstract, title, controlled terms and classification text as content and discarding the rest (e.g. author, publication-year, and language-code). The following preprocessing steps were carried out: normalising all upper case characters to lower case, removing stopwords, and filtering out all terms which occur in more than half of all documents. Stemmed index terms are obtained by applying the German or English Snowball stemmer (provided in the Lucene software) to topics and documents, For the retrieval experiments, the topic title and topic description were used as queries to Lucene.

While the Lucene software provides some support for decompounding in contributed modules, many changes were necessary to achieve the functionality required to conduct experiments on sub-word indexing. Decompounding words into CVC sequences was added as a new tokenizer generating multiple sub-words per word. For CVC sequences and n-grams (and variants), an additional word boundary marker was used (i.e. the character "|") at the beginning and end of a word. Lucene also provides a method to perform dictionary-based decompounding. Preliminary tests indicated that indexing with this method is very time-consuming (and will literally take days) due to inefficient lookup operations in the dictionary. Therefore, the dictionary representation in this method was changed from a set of words to a ternary search tree [Bentley and Sedgewick, 1997], which drastically improves indexing time. German and English (British English spelling) dictionaries were compiled from OpenOffice resources[6]. The German dictionary contains 133,379 entries, the English dictionary contains 46,280. The difference in the number of entries indicates the productivity of the German language to form new words as compounds.

For the hyphenation-based decompounding, hyphenation grammar files for German and English were provided by the Objects For Formatting Objects (OFFO) Sourceforge project.[7] Hyphenation points are inserted into words, defining syllable-like sub-words. Sub-words are required to have a minimum of 2 characters before and 2 characters after a hyphen, i.e. all sub-words have a minimum length of two characters. The character sequences between word boundaries and the hyphenation points are extracted as sub-words.

Time and disk space requirements for indexing and searching were calculated as the average number for two runs. The experiments were performed on a standard PC (Intel Core 2 Duo @ 3 GHz CPU, 4 GB memory, Western Digital 3200AAKS hard disk, OpenSuSe version 10.3 operating system).

[5] http://lucene.apache.org/

[6] http://wiki.services.openoffice.org/wiki/Dictionaries/

[7] http://offo.sourceforge.net/hyphenation/index.html

5 Results and Discussion

Results for the German and English retrieval experiments are shown in Table 2. The following subsections describe retrieval performance, disk and time requirements, and a per-topic analysis of sub-word indexing.

5.1 Retrieval Performance

For German, with the exception of n-grams, all methods for indexing sub-words achieve a higher performance in comparison to stemming. The best performing sub-word indexing methods are to use CVC sequences and index them together with word stems (DE6: +17% MAP, +37% GMAP, +14% rel_ret), or to use syllable-like sub-words obtained from the hyphenation algorithm together with stems (DE12: +9% MAP, +23% GMAP, +11% rel_ret). Figure 3 shows the recall-precision graph for the experiments DE0, DE6, and DE12. The top five methods for German ordered by decreasing MAP are: CVC+stem, VCV+stem, HYPH+stem, DICT+stem, and stem.

An index comprising sub-words in some cases leads to a higher performance (e.g. DE5 vs. DE0, DE7 vs. DE0) compared to the baseline. An index with a combination of stopwords and stems always yields a higher performance compared to indexing sub-words only (e.g. DE2 vs. DE1, DE6 vs. DE5). Both recall (rel_ret) and precision (MAP, GMAP) are improved in the best experiments. In many cases, the number of relevant documents is higher than in the baseline (e.g. DE2, DE5, DE10, DE12). In most experiments, the initial precision (P@10, P@20) does not improve (e.g. DE13-DE18) or does not improve considerably (e.g. DE6 vs. DE0, DE12 vs. DE0).

The dictionary-based decompounding approach was expected to perform worse than approaches not requiring language-dependent or domain-specific resources, because the document corpus has a domain-specific vocabulary. Dictionary-based decompounding performs only slightly better than the baseline (e.g. DE10 vs. DE0).

Hyphenation was expected to outperform overlapping n-grams and CVC sequences, because the results correspond more to meaningful sub-words. Compared to CVC sequences, sub-words spanning word constituents are avoided by hyphenation, i.e. long consonant or vowel sequences spanning constituent words as in *"Geschäftsplan"* (business plan) or *"Seeigel"* (sea urchin) do not occur. Performance of the hyphenation is the second best for all methods (DE12) and clearly outperforms all n-gram methods.

Using overlapping character n-grams as indexing terms does not increase performance (DE13-DE18 and EN13-18). However, no combination of grams with different sizes was tried because combinations of other sub-words were not investigated in this paper (e.g. CV combined VC) and because of the additional disk space requirements.

The MAP for the experiments using CVC (DE6) and hyphenation-based sub-word indexing (DE12) is significantly higher than the MAP for the baseline experiment (Wilcoxon matched-pairs signed-ranks test, N=149, $p <= 0.0001$ and $p <= 0.05$ respectively).

For English, indexing sub-words does not outperform the baseline using standard retrieval on stemmed word forms (EN6: –8% MAP, –11% GMAP, +1% rel_ret for using CVC and stems). For two experiments, indexing CVC

```
<DOC>
 <DOCID> GIRT-EN19900121783 </DOCID>
 <TITLE> Measures and projects of the Land of Lower Saxony for
  combatting female unemplyoment </TITLE>
 <AUTHOR> Wigbers, Antonia </AUTHOR>
 <PUBLICATION-YEAR> 1989 </PUBLICATION-YEAR>
 <LANGUAGE-CODE> EN </LANGUAGE-CODE>
 <COUNTRY-CODE> DEU </COUNTRY-CODE>
 <CONTROLLED-TERM> Lower Saxony </CONTROLLED-TERM>
 <CONTROLLED-TERM> woman </CONTROLLED-TERM>
 <CONTROLLED-TERM> employment promotion </CONTROLLED-TERM>
 <CONTROLLED-TERM> unemployment </CONTROLLED-TERM>
...
 <METHOD-TERM> documentation </METHOD-TERM>
 <METHOD-TERM> applied research </METHOD-TERM>
 <CLASSIFICATION-TEXT> Employment Research </CLASSIFICATION-TEXT>
</DOC>
```

Figure 1: Sample English GIRT4 document.

```
<top>
 <num> 177 </num>
 <EN-title> Unemployed youths without vocational training </EN-title>
 <EN-desc> Find publications focusing on jobless adolescents who have
  not completed any vocational training. </EN-desc>
 <EN-narr> Relevant documents give an overview of the scale and the
  problem of jobless adolescents who have not completed any job training.
  Not relevant are documents dealing exclusively with measures for youth
  welfare and youth policy. </EN-narr>
</top>
```

Figure 2: Sample English GIRT4 topic.

and hyphenated sub-words together with stems, the number of relevant and retrieved documents is slightly higher than in the baseline experiment (EN6 and EN12). No experiment improved MAP, GMAP or the precision at N documents in comparison to the baseline. The top five methods for English are: stem, CVC+stem, DICT+stem, HYPH+stem, and VCV+stem.

5.2 Disk space and time requirements

In addition to traditional information retrieval performance, the requirements to index and search the document collection using sub-word indexing were measured. More complex preprocessing requires more time, i.e. the time needed to process documents and queries increases.

All methods using a combination of stems and sub-words as indexing terms need more time and space than the baseline, which was an expected outcome. In the index combining stems and sub-words, all indexing terms from the baseline (stems) have to be generated and stored in addition to the sub-words. Dictionary-based decompounding requires the most additional time for indexing the document collection (+225.2% increase compared to the baseline). Hyphenation-based decompounding requires the most additional time for searching (+364.1%). However, a longer processing time is no guarantee for a better performance, as is shown by the dictionary-based approach.

5.3 Topic Analysis

The best two methods for decompounding (DE6 and DE12) were analysed in more detail on a per-topic basis. To obtain the average number of compounds in the topics, the following rules and guidelines for counting compound words were established:

- Abbreviated coordinations with hyphens count as one compound (e.g. *"Parlaments- oder Präsidentschaftswahlen"*).
- Words with bound morphemes count as a compound (e.g. *"Kneipengänger"*).
- Words with non-separable prefixes count as a compound (e.g. *"Ökosteuer"*).
- Hyphenated words do not count as compound words (e.g. *"burn-out"*).
- Compounds are not limited to nouns, but also include verbs and adjectives (e.g. *"rechtsextrem"*).
- Words which may be incorrectly decomposed into constituent words do not count as compounds (e.g. *"Mutterschaft"*).

Following these guidelines, the GIRT topics were manually annotated. The 150 topics contain an average of 1.49 compounds per topic. The topics for the best-performing methods were sorted by gain in MAP. For CVC (DE6), the average number of compounds in the top-20 topics is 2.15, for HYPH (DE12) the average is 2.3. There is an overlap of 17 topics of the top-20 best performing topics for experiments DE6 and DE12.

There are two topics among the top-20 which do not contain any compounds at all, topic 82: *"Berufliche Bildung von Immigranten"* (Professional training of immigrants)/ *"Finde Dokumente, die über die berufliche Integration von Immigranten durch berufliche Bildung berichten"* (Find

Table 2: Results for monolingual retrieval experiments on German and English GIRT4 documents (lang.: language; rel_ret: number of relevant and retrieved documents).

Run	Parameters	Results							
ID	index terms	rel_ret	MAP	GMAP	P@10	P@20	indexing [s]	searching [s]	index size [MB]
DE0	stem	11025	0.3214	0.2097	0.63	0.55	279.5	**17.3**	**659**
DE1	CV	10778	0.2715	0.1554	0.52	0.46	338.8 (+21.2%)	32.5 (+87.8%)	1106 (+67.8%)
DE2	CV+stem	12108	0.3494	0.2537	0.62	0.55	576.6 (+106.2%)	40.9 (+136.4%)	1412 (+114.2%)
DE3	VC	10480	0.2399	0.1308	0.47	0.41	339.7 (+21.5%)	68.4 (+295.3%)	1075 (+63.1%)
DE4	VC+stem	11819	0.3317	0.2448	0.60	0.54	532.5 (+90.5%)	43.3 (+150.2%)	1383 (+109.8%)
DE5	CVC	12360	0.3584	0.2673	0.63	0.56	472.6 (+69.0%)	52.7 (+204.6%)	1285 (+94.9%)
DE6	CVC+stem	**12599**	**0.3765**	**0.2886**	**0.65**	**0.58**	631.8 (+126.0%)	39.0 (+125.4%)	1585 (+140.5%)
DE7	VCV	11879	0.3311	0.2309	0.59	0.53	358.7 (+28.3%)	37.2 (+115.0%)	1185 (+79.8%)
DE8	VCV+stem	12477	0.3654	0.2771	0.63	0.56	729.5 (+161.-%)	49.6 (+186.7%)	1492 (+126.4%)
DE9	DICT	11545	0.3051	0.1958	0.53	0.49	617.2 (+120.8%)	63.8 (+268.7%)	1170 (+77.5%)
DE10	DICT+stem	12252	0.3450	0.2447	0.61	0.53	909.2 (+225.2%)	75.0 (+333.5%)	1376 (+108.8)
DE11	HYPH	11743	0.3217	0.2269	0.59	0.53	433.5 (+55.0%)	40.4 (+133.5%)	896 (+35.9%)
DE12	HYPH+stem	12291	0.3511	0.2582	0.62	0.56	682.0 (+144.0%)	80.3 (+364.1%)	1111 (+68.5%)
DE13	3-gram	10380	0.2518	0.1546	0.51	0.45	473.6 (+69.4%)	67.3 (+289.0%)	1582 (+140.0%)
DE14	3-gram+stem	10901	0.2835	0.1940	0.54	0.50	774.7 (+177.1%)	70.8 (+309.2%)	1809 (+174.5%)
DE15	4-gram	9961	0.2429	0.1590	0.52	0.47	376.3 (+34.6%)	51.1 (+195.3%)	1338 (+103.0%)
DE16	4-gram+stem	10180	0.2547	0.1716	0.54	0.48	633.8 (+126.7%)	54.2 (+213.2%)	1503 (+128.0%)
DE17	5-gram	7824	0.1765	0.0911	0.48	0.41	**277.5 (-0.8%)**	29.5 (+70.5%)	964 (+46.2%)
DE18	5-gram+stem	8095	0.1876	0.1017	0.50	0.43	352.5 (+26.1%)	48.1 (+178.3%)	1058 (+60,.5%)
EN0	stem	**10911**	**0.3453**	**0.2239**	**0.57**	**0.53**	179.6	**12.0**	**275**
EN1	CV	9027	0.2144	0.1049	0.43	0.38	171.3 (-4.7%)	25.0 (+108.3%)	493 (+79.2%)
EN2	CV+stem	10573	0.3002	0.1804	0.54	0.48	268.9 (+49.7%)	32.0 (+166.6%)	626 (+127.6%)
EN3	VC	8576	0.1800	0.0797	0.38	0.34	174.5 (-2.9%)	23.8 (+98.3%)	483 (+75.6%)
EN4	VC+stem	10551	0.2953	0.1802	0.54	0.48	265.2 (+47.6%)	29.4 (+145.0%)	615 (+123.6%)
EN5	CVC	10545	0.2929	0.1775	0.55	0.48	186.9 (+4.0%)	25.9 (+115.8%)	551 (+100.3%)
EN6	CVC+stem	10985	0.3181	0.1993	0.56	0.50	304.8 (+69.7%)	30.9 (+157.5%)	679 (+146.9%)
EN7	VCV	10082	0.2649	0.1557	0.51	0.45	189.0 (+5.2%)	30.8 (+156.6%)	526 (+91.2%)
EN8	VCV+stem	10759	0.3074	0.1952	0.56	0.50	255.6 (+42.3%)	30.1 (+150.8%)	658 (+139.2%)
EN9	DICT	10163	0.2797	0.1587	0.53	0.47	281.9 (+56.9%)	38.1 (+217.5%)	561 (+104.0%)
EN10	DICT+stem	10785	0.3139	0.1915	0.55	0.50	390.7 (+117.5%)	41.9 (+249.1%)	640 (+132.7%)
EN11	HYPH	10451	0.2813	0.1740	0.53	0.46	206.4 (+114.9%)	23.4 (+95.0%)	376 (+36.7%)
EN12	HYPH+stem	10908	0.3104	0.1944	0.53	0.48	303.7 (+69.0%)	28.1 (+134.1%)	460 (+67.2%)
EN13	3-gram	9549	0.2388	0.1410	0.49	0.43	228.3 (+27.1%)	43.9 (+265.8%)	712 (+158.9%)
EN14	3-gram+stem	9989	0.2678	0.1668	0.53	0.47	295.3 (+64.4%)	48.3 (+302.5%)	831 (+202.1%)
EN15	4-gram	8709	0.2149	0.1128	0.47	0.41	173.6 (-3.4%)	22.2 (+85.0%)	573 (108.3%)
EN16	4-gram+stem	8964	0.2317	0.1238	0.50	0.44	260.6 (+45.1%)	27.6 (+130.0%)	663 (+141.0%)
EN17	5-gram	6236	0.1482	0.0611	0.42	0.35	**146.2 (-18.6%)**	15.4 (+28.3%)	388 (+41.0%)
EN18	5-gram+stem	6354	0.1535	0.0660	0.43	0.36	207.6 (+15.5%)	16.0 (+33.3%)	439 (+59.6%)

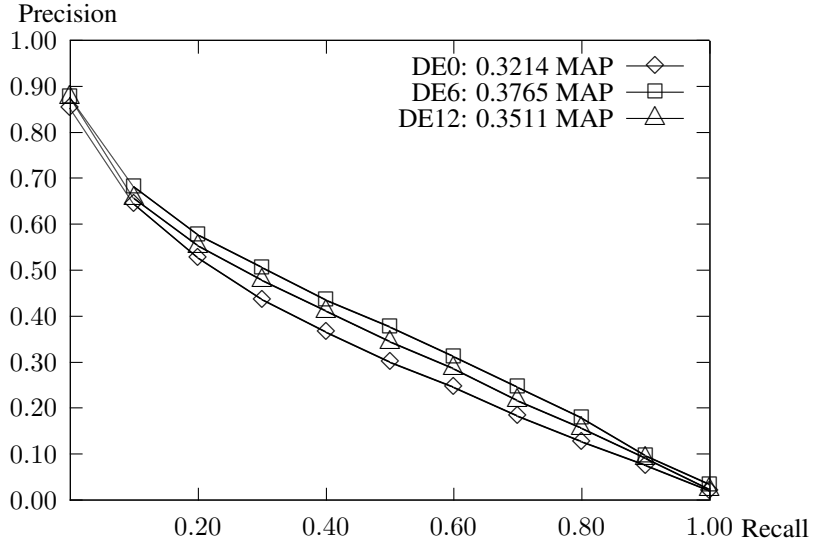

Figure 3: Recall-precision graph for selected experiments.

documents on the professional integration of immigrants through vocational training) and topic 101: *"Tiere in der Therapie"* (Animals in therapy)/ *"Finde Dokumente, die über das Nutzen des Potenzials von Tieren in der therapeutischen Arbeit mit dem Menschen berichten"* (Find documents reporting on the potential of using animals in human therapeutic programs).

In topic 82, standard IR methods like stemming do not allow matching possibly relevant documents mentioning *"Immigration"* (immigration) instead of *"Immigranten"* (immigrants). If decompounding is used to split words into sub-words, these different but semantically related words will have some sub-words in common and additional documents can be found.

For topic 101, the terms *"Therapie"* (therapy) and *"therapeutisch"* (therapeutical) are usually stemmed to different indexing terms and each will have a low weight assigned to them. Using sub-words, these word forms share some sub-words assigned to them and the shared sub-words will have a higher weight. In addition, this topic contains a word with the new German spelling, *"Potenzial"* (potential)). Most documents in the GIRT collection were written before the spelling was changed. The term *"Potential"* in the old German spelling has a term frequency of 2323, *"Potenzial"* has a frequency of 76. Thus, very few documents containing the new spelling will be found. Matching terms on a sub-word level (instead of exact matching on the word-level) will yield more potentially relevant documents.

5.4 Summary

In summary, sub-word indexing does not perform equally for the non-compounding language English in comparison to the compounding language German. Most German experiments clearly outperform the stemming baseline with respect to retrieval metrics MAP, GMAP, P@10, and P@20.

All sub-word indexing methods require more time for indexing and searching a database. In addition, the index size for sub-words is higher compared to a stem index. The size of a combined index (using sub-words and stems as indexing units) is up to an additional 174% of the original size.

Indexing time for 5-grams is lower than the indexing time for the stemming baseline. The required time to index and search a collection increases with the number of indexing units produced. In a combined index (sub-words and stems), the stems also have to be produced. Additionally, typically several sub-words are identified for each word. Thus, indexing and searching sub-words requires more time than for the stemming baseline.

The best performing methods – CVC indexing and hyphenation-based sub-word indexing – perform significantly better than the stemming baseline for German, they perform best on very similar topics, and they even improve some topics which do not contain compounds at all.

6 Conclusion and Future Work

Four different approaches to break up words for indexing sub-words were discussed and evaluated on the German and English data for the domain-specific track GIRT at CLEF. Three of the methods outperform the stemming baseline. These methods include consonant-vowel sequences, which have been mostly used for spoken document retrieval and a new method for decompounding, based on hyphenation patterns to find sub-words. In comparison to the standard stemming baseline, decompounding yields a significantly higher performance in terms of MAP, GMAP, and rel_ret for German. In conclusion, sub-word indexing for German may be seen as a method integrating decompounding and stemming: words are broken down into smaller indexing units and frequent affixes are either removed completely or are associated with a low weight.

The best performing methods are also very cost-effective and easily adaptable to other languages. Consonant-vowel sequences can be produced as a by-product of stemming and stemmers already exist for many languages. Snowball contains stemmers for about 16 languages. Similarly, there already are TeX hyphenation rules for more than 30 different languages as well. Indexing n-grams did not produce results comparable to or higher than the stemming baseline. For English, sub-word indexing does not perform as good as stemming, most likely because English words do not have to be split into smaller units.

Splitting compounds into several smaller indexing units

considerably changes many implicit parameters for IR, including the number of terms in both queries and documents, term frequencies, and the average document length. These changes suggest that parameters should be adjusted and optimised correspondingly if a different weighting model is applied. Future work will include experiments with state-of-the-art retrieval models (e.g. OKAPI BM25, [Robertson *et al.*, 1994]), determining parameters based on the new characteristics of the index and topics. The effect of sub-words on relevance feedback will be investigated for different sub-word indexing methods.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project.

References

[Bentley and Sedgewick, 1997] Jon L. Bentley and Robert Sedgewick. Fast algorithms for sorting and searching strings. In *SODA '97: Proceedings of the eighth annual ACM-SIAM symposium on discrete algorithms*, pages 360–369, Philadelphia, PA, USA, 1997. Society for Industrial and Applied Mathematics.

[Braschler and Ripplinger, 2003] Martin Braschler and Bärbel Ripplinger. Stemming and decompounding for German text retrieval. In F. Sebastiani, editor, *ECIR 2003*, volume 2633 of *Lecture Notes in Computer Science (LNCS)*, pages 177–192, Berlin, 2003. Springer.

[Chen and Gey, 2004] Aitao Chen and Fredric C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7(1–2):149–182, 2004.

[Chen, 2003] Aitao Chen. Cross-language retrieval experiments at CLEF 2002. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19-20, 2002. Revised Papers*, volume 2785 of *Lecture Notes in Computer Science (LNCS)*, pages 28–48. Springer, Berlin, 2003.

[Daumke *et al.*, 2007] Philipp Daumke, Jan Paetzold, and Kornel Marko. MorphoSaurus in ImageCLEF 2006: The effect of subwords on biomedical IR. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, Revised Selected Papers*, volume 4730 of *Lecture Notes in Computer Science (LNCS)*, pages 652–659. Springer, Berlin, 2007.

[Daumke, 2007] Philipp Daumke. *Das MorphoSaurus-System – Lösungen fur die linguistischen Herausforderungen des Information Retrieval in der Medizin*. PhD thesis, Albert-Ludwigs-Universität, Freiburg i.Br., Medizinische Fakultät, 2007.

[Glavitsch and Schäuble, 1992] Ulrike Glavitsch and Peter Schäuble. A system for retrieving speech documents. In *Proceedings of ACM SIGIR 1992*, pages 168–176, Denmark, 1992.

[Kamps *et al.*, 2003] Jaap Kamps, Christof Monz, and Maarten de Rijke. Combining evidence for cross-language information retrieval. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19-20, 2002. Revised Papers*, volume 2785 of *Lecture Notes in Computer Science (LNCS)*, pages 111–126. Springer, Berlin, 2003.

[Kluck, 2005] Michael Kluck. The domain-specific track in CLEF 2004: Overview of the results and remarks on the assessment process. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pages 260–270. Springer, Berlin, 2005.

[Knuth, 1984] Donald E. Knuth. *Computers & Typesetting. Volume A. The TeXbook*. Addison-Wesley, Reading, Mass., 1984.

[Liang, 1983] Franklin Mark Liang. *Word hy-phen-a-tion by com-put-er*. PhD thesis, Stanford University, Department of computer science, Stanford, CA, USA, 1983.

[McNamee and Mayfield, 2007] Paul McNamee and James Mayfield. N-gram morphemes for retrieval. In *Working Notes of the CLEF 2007 Workshop*, Budapest, Hungary, September 2007.

[McNamee, 2001] Paul McNamee. Knowledge-light Asian language text retrieval at the NTCIR-3 workshop. In Keizo Oyama, Emi Ishida, and Noriko Kando, editors, *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, Tokyo, Japan, 2001. National Institute of Informatics (NII).

[Ng, 2000] Kenney Ng. *Subword-based approaches for spoken document retrieval*. PhD thesis, Massachusetts institute of technology (MIT), Department of electrical engineering and computer science, 2000.

[Porter, 1980] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[Robertson *et al.*, 1994] Stephen E. Robertson, Steve Walker, Susan Jones, and Micheline Hancock-Beaulieu. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)*, Gaithersburg, USA, 1994.

[Schäuble and Glavitsch, 1994] Peter Schäuble and Ulrike Glavitsch. Assessing the retrieval effectiveness of a speech retrieval system by simulating recognition errors. In *HLT '94: Proceedings of the workshop on Human Language Technology*, pages 370–372, Morristown, NJ, USA, 1994. Association for Computational Linguistics.