Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-111

Managing Word Form Variation of Text Retrieval in Practice – why Five Character Truncation Takes it all?

Kimmo KETTUNEN[1]
School of Information Sciences, University of Tampere, Finland

Abstract. This paper discusses different methods that have been used for management of word form variation in information retrieval during the history of textual information retrieval. The techniques have been characterized in many ways during the history of IR. We pinpoint the most meaningful features of the approaches and make comparisons that have practical value. In the discussion we characterize word form variation management methods in different ways and offer the reader an overall practical guide for choosing between different methods to be used.

Keywords. information retrieval, management of word form variation, comparison of word form variation management methods, IR performance, effectiveness

Introduction

One of the basic problems of full-text retrieval is variation of word forms that is caused by morphology of natural languages. Shortly put, this means that one base or dictionary form of a word in language may occur in different (inflected) variant forms in texts. Out of this follows that many times the principle "one keyword – one concept - one match" does not hold in the textual index of retrieval systems due to morphology alone. Consequently something needs to be done to morphological variation so that the performance of information retrieval (IR) systems will not suffer too much if the language has a rich or at least medium rich morphology.

To overcome the problem of keyword variation several management methods have been proposed during the history of textual IR. The first word analysing method applied to IR was stemming, first stemmer being Janet Lovins's stemmer for English [1]. Late 1980's saw the in-march of morphological analysis using large dictionaries, also known as lemmatization [2, 3]. During the last 10 years unsupervised morpheme detection methods [4] have been used somehow successfully in management of word form variation management of IR [5]. All these methods can be characterized as

[1] Corresponding author: School of Information Sciences, Information Studies and Interactive Media, FI-33014 University of Tampere, Finland; E-mail: kkettun4@welho.com

reductive [6]: running word forms are analyzed in them and reduced to either stems or base forms or morphs, if possible. The reduced forms are then used both in the indexes of search engines and as keywords in searches.

Another logical option for management of keyword variation is to use generated inflected word forms (or only inflectional stems) as search keys. In this approach, a set of inflected variant forms is generated from the input keyword and these are sought for in the plain word index of the retrieval engine. The basic fear in this method is that the language has too much inflection and too many generated word forms need to be sought for, which would make search impractical due to time considerations. But as e.g. Kettunen and Airio [7], Kettunen and Arvola [8] and Leturia et al. [9] have shown, only a partial generation of the most frequent inflected word forms yields good retrieval performance.

So far mentioned methods can be characterized as linguistically motivated, either fully (morphological analysis, word form generation) or partly (stemming, unsupervised morpheme detection). A third group of methods is non-linguistic, and it includes different character string oriented methods. These include truncation of keywords, character n-gramming [10] and usage of hyphen like structures [11]. Truncation was perhaps the first method of word form variation management used in IR, and it was first based on the user's choice of proper truncation point. Lately, truncation with a fixed length (e.g. five character truncation starting from the beginning of the word) has been shown quite effective with many languages [10]. These methods and their variants cover most of the word form variation management techniques that are actively used in IR.

The structure of the article is following: we shall first give a short account of information retrieval basics and after that we proceed to discuss different word form management methods and their relative advantages. Our basic findings and recommendations are all in Chapter 2, and discussion draws some more conclusions on the issue.

1. IR basics

For our discussion we need to first outline working principles of a state-of-the-art text search engine. Our description is based on two current textbooks, Croft et al. [12] and Ingwersen and Järvelin [13]. Due to space requirements our discussion is very concise, and an interested reader is asked to look after the references for further details.

By a text information retrieval system we mean a textual database system consisting of text documents and means to manage the database. Documents in the database can be searched for, and new documents can be added to the database if needed. Searching in the textual database is based on matching of a query term representation and an inverted index that represents the contents of the documents as index terms. Matching of search keys or terms can be either full or partial. Full matching IR systems are Boolean, partial matching systems can be e.g. statistical. Most of the modern search engines are partial match systems. A partial match (or best-match) IR system does not require an exact match of the query and document terms, and thus it is able to return documents that match the query only partially. Another important feature of partial match IR systems is *ranking*: returned documents are given as an ordered list where the documents expected to be most relevant are at the top and less relevant in decreasing order of relevance. [12, 13-28; 13, 119]. Figure 1 gives an outline of the overall situation including the search engine user.

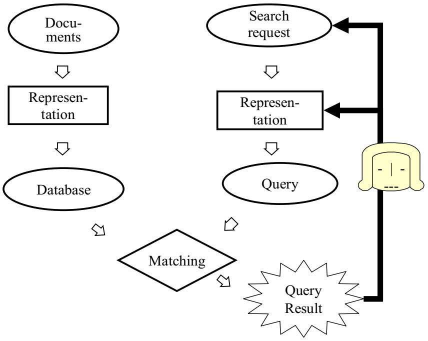

Figure 1. A simplified picture of an IR system, adapted from Ingwersen and Järvelin [13, 115]

The basic goal for an IR engine is to fulfil user's information need as well as possible. The more the engine returns relevant documents at the top of the result list, the better. Users, however, may be satisfied with only a few highly relevant documents at the beginning of the result list. This is especially true with web searches [13]. Management of word form variation in an IR engine *may* help in achieving this goal.

2. When should word form variation management be used?

As shown in the IR basics part, queries and documents are matched in the IR database according to their string level token representations. Singular and plural surface forms of lexeme {*cat*}, *cat* and *cats*, do not match, if something, like stemming, is not done to them to make the representations similar. In all of the word form variation management methods the basic principle is the same: decreasing of variation found in natural language word forms.

When one reads IR literature, it is easy to notify, that word form variation management is used many times with a language that would not actually need any word form variation management, because the language is morphologically so simple [14]. During most of the 1970s and 1980s different stemming algorithms were evaluated with different English IR collections, as no other collections were usually available. In retrospect, at least part of this work seems futile: there just is not much to be gained with IR performance of English. The same holds for many other languages, too.

The explanation for varying behavior of words in different languages is linguistic complexity. On morphological level linguistic complexity means roughly, that the language has lots of inflection, which is realized, for example, in number of different nominal case forms the language has (e.g. Iggesen [15]). Finnish, for example, has 14 different cases, and English has two. This means that Finnish has many variating word forms, as English has few. Most of the European languages fall between these figures. Many times already the number of cases in the language is indicative of morphological complexity, but not always (e.g. in Swedish and in Bulgarian). Then other morphological categories, such as marking of definiteness and expression of number in the language or compounding, are the key factors.

The ‘IR hardness’ of a language is clearly related to its linguistic complexity. McNamee et al. [10, Table 6] show this by relating length of words (the longer the words in the language, the more morphemes they have), two linguistic complexity ratios and gains in IR performance achieved with 5-grams. These figures correlate at least moderately (lowest correlation being 0.68) or very highly (highest correlation being 0.91). Kettunen [6] shows the same with counting the difference of best and worst mean average precision (MAP) results of IR performance for the language. The bigger the difference, the more morphologically complex the language is.

For further demonstration and discussion of the importance of this point, we will proceed with some data from two different studies that have either empirical results from IR evaluations of several languages or have collected such data from other studies.

McNamee et al. [10] use 18 different methods for management of word form variation for 18 languages in five different writing systems. Methods for word form variation management include all the main methods used in IR, except lemmatization that is not easily available for such a variety of languages. Instead of lemmatization, two types of stemmers, rule-based and statistical, are used. Different phonetic transformations (soundex and devowelization), truncations and character gramming (n-gramming, where n varies from 3 to 7, and skip-gramming, where some of the characters may be skipped) are included in the methods. The main results of the paper are the following:

- character n-gramming is the most effective method for most of the languages
- rule based stemming (Snowball stemmers are used) can be an attractive option for languages where morphological variation is not very high
- phonetic transformations do not work well for any language
- a statistical stemmer (i.e. particular unsupervised morphological method) does not perform too well, but is getting better (cf. also Kurimo et al. [5] for the latest results with different systems)
- one of the most unsophisticated and un-linguistic methods, five character truncation, works very well with most of the languages, being the second best non n-gram method overall, only slightly behind performance of Snowball stemmers.

Table 1 combines results of McNamee et al. [10] and collected IR data from Kettunen [6], and shows the situation with 14 languages that have available IR collections and data. Many interesting small languages, such as Estonian, Latvian and Lithuanian, are unfortunately missing from the table, as there are no IR collections for these languages, but the variation in languages is enough to make our points.

Columns two and three in the table show basically the same thing, difference between the IR result when best possible word form variation management method has been used for the language versus situation when plain word forms have been used in queries. Column four interprets need for word form variation management according to Sparck-Jones's [16] old rule: if the statistically significant absolute difference in MAP is under 5 %, the practical difference is not noticeable; if the MAP difference is over 5 % but under 10 %, the practical difference is noticeable. When the difference is over 10 %, the practical difference is material. These figures are stated as *no need*, *beneficial* and *necessary* in the Table.

Table 1. Necessity of word form variation management in the light of MAP results

Language	GAP = best MAP with word form variation management *minus* plain words MAP [6]	Lowest and highest MAPs gained [10] *low*	 *high*	Is word form variation management needed for the language?
1. Bulgarian	6.8-8.1 %	0.216	0.31	beneficial
2. Czech	N/A	0.227	0.329	necessary
3. Dutch	0.6.-5.0 %	0.381	0.424	no need
4. English	1.2-2.9 %	0.406	0.437	no need
5. Finnish	10.5-25.2 %	0.34	0.507	necessary
6. French	0.5-3.8 %	0.363	0.401	no need
7. German	6-15.7 %	0.33	0.42	beneficial/necessary
8. Hungarian	9.9-12.4 %	0.197	0.374	necessary
9. Italian	N/A	0.374	0.417	no need
10. Portuguese	N/A	0.316	0.352	no need
11. Russian	6.1-21.0 %	0.267	0.373	necessary
12. Spanish	N/A	0.439	0.484	no need/beneficiary
13. Swedish	1.7-8.8 %	0.338	0.427	beneficial
14. Turkish	12.3 %	N/A[2]		necessary

In some cases (Bulgarian, German and Swedish), the line between *beneficial* and *necessary* is quite narrow, and in most of the cases of *no need,* there is no question of the borderline. Only Spanish seems to be not too far from the 5 per cent edge.

2.1. Other criteria

Many methods of word form variation management of IR work considerably well from the viewpoint of effectiveness, which is measured in precision and recall (P/R) of retrieval using different measures, one of the most used being MAP in Table 1. The methods can also be compared on a more general level. Three kinds of benefits are usually associated with different types of keyword variation management in IR [17].

[2] McNamee et al. do not have Turkish in their repertoire, but empirical results of Kettunen et al. [11] confirm GAP result shown in Kettunen [6] with several word form management methods.

They are briefly as follows:

- ease of use (morphology of query words is taken care of by the retrieval system),
- storage savings - the index compression factor, ie. smaller indexes when for example lemmatization or stemming is used [14], and
- improved retrieval performance.

Besides these criteria, there are, however, others that should be taken into consideration. Linguistic methods of word form variation management use many times lexicons in their analysis, and thus the lexical coverage of the morphological method used is important. This is an issue that affects lemmatizers and stemmers using dictionaries. Their dictionaries lack words for many reasons, and one of the main classes of lacking words are different kinds of proper names (persons, companies, geographical names etc.), which are usually an important subclass of query words [18]. A statistical lemmatizer, such as e.g. Stale [19], in turn, does not suffer from this hinder, and performs also competitively with a lexical lemmatizer in an IR context. Word form generators can be implemented without lexicons, and thus they avoid the problem of lexical coverage.

Other more technical criteria can also be used for comparison. Croft et al. [12, 327] list the following: elapsed indexing time, indexing processor time, query throughput, query latency, indexing temporary space and index size. These criteria are related to search engine efficiency and are especially important when commercial search engines are developed and used.

We have chosen to Table 2 five different evaluation criteria for word form variation management methods used in IR. The criteria are language independence of the method, its IR effectiveness, size of the retrieval indexes created with the method, possibility of automatic generation of the rules of the management method and overall simplicity of the approach. These criteria are by no means exhaustive and also others could be included or some omitted. Efficiency considerations have been left out of our criteria, because there is not available data related to them and efficiency is also so dependent on specific implementation.[3]

The methods have been assessed with points 0, +2 and +4. With zero the effect is not positive or not applicable, with +2 effect is clearly positive, mid-size, and with +4 there is a big positive effect, best performance. Two different experienced IR researchers besides the author gave points to the methods. Figures given in the Table 2 are means from these three assessments except in case of unsupervised morphological methods, where only two assessments were given.

When figures of the Table 2 are examined, we can see that character oriented methods get the best points. Five character truncation, unsupervised morphological methods and syllabification are the three best methods here, in this order. Rule-based stemming and lemmatization with rules and a dictionary do not fare too well, although they are the two most used methods of word form variation management in IR research. The results and the chosen assessment criteria are of course open to discussion, but in our opinion they do reflect important details that should be taken into consideration when choosing word form variation management methods for an IR system.

[3] Harman's [17] ease of use is omitted from the table, as it is actually included in all of the methods.

Table 2. Scoring of different word form variation management methods along five criteria

Method	Language independence	Effective-ness	Index size	Automatic generation of rules	Simplicity of the approach	SUM
automatic truncation [10]	4	3.33	3.33	2.66	4	**17.32**
unsupervised morphological methods [4,5]	4	4	3	2	2	**15.0**
syllabification [11]	3.33	2.66	2	2.66	3.33	**13.98**
n-gramming (plain, no skips) [10]	4	3.33	0	2	3.33	**12.66**
statistical lemmatization [19]	2.66	3.33	1.33	2	1.33	**10.65**
rule based stemming [10]	0.66	2.66	3.33	0.66	2	**9.31**
plain words	4	0	1.33	0	3.33	**8.66**
word form generation [6]	0.66	4	1.33	1.33	1.33	**8.65**
lemmatization (rules + dict.) [3]	0	4	2	0	0.66	**6.66**

2.2. A heuristics for use of word form variation management methods

Based on the data in Tables 1 and 2 a heuristic recommendation for usage of different word form variation management methods in IR would be like this – the heuristics applies for other languages not shown here, too.

1. For morphologically simple languages (such as 3, 4, 6, 9 in Table 1) do nothing but normal routines (case folding etc.). Plain word forms are a good solution for indexing and query formation with these languages.
2. If the language is in the *beneficial* group (such as 1, 7, and 13 in Table 1), the simplest non-linguistic word form management method can be used. Out of the simple methods five character truncation is the easiest to implement and very effective, but also n-gramming and hyphenation could be used. Large indexes and slow retrieval are shortcomings of n-gramming. A light stemmer can also be considered, if such is available. But there is no need for 'heavy artillery' here.
3. With languages in the *necessary* group *(*such as 2, 5, 8, 11 and 14 in Table 1) one can begin to consider also 'heavier' methods, such as stemming or lemmatization. Even here they are not necessary, as five character truncation is effective with these languages too. If one's only need is to have the best IR performance from the search engine, then language technology oriented tools may be overkill. If one has also other needs for the linguistic analysis capabilities of the IR system (such as handling of lemmas or interaction as e.g. in query expansion, cf. Galvez et al. [15], then one may consider an elaborate lemmatizer.

3. Discussion

When word form variation management methods of IR are discussed, one needs to keep in mind, that the issue has two dimensions: that of language technology or linguistic processing and that of information retrieval. Language technology and information retrieval have partly different and partly overlapping criteria for developing and using word form handling tools. Language technology aims at linguistic felicity and as broad linguistic coverage as possible [3, 14]. These are justified aims, but they should be kept separate from IR performance the methods enhance. Information retrieval can have more modest aims with its LT tools: it may be satisfied with linguistically poorer methods that improve effectiveness of searches in its current phase, where retrieval is based on matching of string level representations of words, not semantic entities.

Our considerations and suggestions come near to Church's DDI claim (Don't Do It), which states that morphology aware software should perhaps not be used at all in computational handling of language: "*There are lots of morphology programs out there, many of which work surprisingly well. Nevertheless, for many practical applications, we prefer not to use such programs, if we have the choice. Simple morphological inferences are better than complex inferences. But even simple inferences are worse than none*." [20] When examining our data, this seems partly true with regards of the role of morphology programs in IR: you can skip proper morphological processing with use of simple string manipulation and get good results anyhow. Some time all morphological inferences can be skipped (cf. Table 1, no need languages), and most of the times simple inferences do the trick.

Another, more theoretical, argument in favor of simpler methods is Minimal Description Length (MDL), which basically formalizes the old Occam's razor: when two models fit the data equally well, MDL will choose the one that is the simplest in the sense that it allows for a shorter description of the data [21, 29-]. If we apply the idea of MDL for morphological components used in IR, we can e.g. say that five character truncation could be favored instead of a lemmatizer, as it is far simpler and "fits the data" – i.e. management of word form variation for IR – as well as stemming or lemmatization with many languages. A five character truncation module for a search engine can be coded in about two to three code lines in almost any programming language, when a lexical lemmatizer needs description of lexicons (tens of thousands of lines) and a rule component (a few hundred lines). The same argument would apply for simple syllabification, although it is slightly more complex on the index side representation. Other methods are between these extremes.

We have discussed usefulness of different word form variation management methods and given some practical hints for choosing the methods for IR purposes. The issue is far from simple, and many arguments can be given pro different solutions. We have taken a low-level approach, where need of very high level morphological tools with IR has been questioned. Perhaps less is really more in this case?

References

[1] J. B Lovins, Development of a Stemming Algorithm, *Mechanical Translation and Computational Linguistics,* 11 (1968), 23–31.

[2] R. Alkula, From plain character strings to meaningful words: producing better full text databases for inflectional and compounding languages with morphological analysis software, *Information Retrieval* 4 (2001), 195−208.

[3] K. Koskenniemi, Finite state morphology and information retrieval, *Natural Language Engineering* 2 (1996), 331–336.

[4] H. Hammarström and L. Borin, Unsupervised Learning of Morphology, *Computational Linguistics* 37 (2011), 309–350.

[5] M. Kurimo, S. Virpioja and V. Turunen, (eds.), Proceedings of the Morpho Challenge 2010 workshop. Technical Report TKK-ICS-R37, Aalto University School of Science and Technology, Department of Information and Computer Science, Espoo, Finland. http://research.ics.aalto.fi/events/morphochallenge2010/papers/ProcMorphoChallenge2010.pdf

[6] K. Kettunen, Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval, *Journal of Documentation* 65 (2009), 267–290.

[7] K. Kettunen and E. Airio, Is a morphologically complex language really that complex in full-text retrieval? In Salakoski, T. et al. (eds.), Advances in Natural Language Processing, LNAI 4139, Springer-Verlag, Berlin Heidelberg, (2006)., 411–422.

[8] K. Kettunen and P. Arvola, Generating variant keyword forms for a morphologically complex language leads to successful information retrieval with Finnish. In Larsen, B. and Salampasis, M. (eds.), Advances in Multidisciplinary Retrieval, 5th Information Retrieval Facility Conference, (2012), 113-126.

[9] I. Leturia, A. Gurrutxaga, N. Areta, I. Alegria and A. Ezeiza, EusBila, a search service designed for the agglutinative nature of Basque. In Lazarinis, F., Vilares, J. and Tait, J.I. (eds.), First workshop on improving non-English web searching (ACM Sigir 2007 Workshop), (2007), 47–54.

[10] P. McNamee, C. Nicholas, C. and K. Mayfield, J., Addressing morphological variation in alphabetic languages, In: Proceedings of the 32nd Annual International Conference on Research and Development in Information Retrieval (SIGIR-2009), Boston, MA, (2009), 75-82.

[11] K. Kettunen, P. McNamee and F. Baskaya, Using syllables as indexing terms in full-text retrieval. In I. Skadina, A. Vasiljevs (eds), Human Language Technologies, the Baltic Perspective, (2010), 225−232.

[12] B. W. Croft, D. Metzler and T. Strohman, *Search Engines. Information Retrieval in Practice.* Pearson, 2010.

[13] P. Ingwersen and K. Järvelin, *The Turn. Integration of Information Seeking and Retrieval in Context.* Springer, 2005.

[14] C. Galvez, F. de Moya-Anegón, and V. H. Solana, Term conflation methods in information retrieval. Non-linguistic and linguistic approaches, *Journal of Documentation* 61(2005), 520–547.

[15] O.A. Iggesen, Number of Cases. In: Dryer, M. S. and Haspelmath, M. (eds.) The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 49A. Available online at http://wals.info/chapter/49A, (2011).

[16] K. Sparck-Jones, Automatic indexing, *Journal of Documentation*, 30 (1974), 393-432.

[17] D. Harman, How effective is suffixing? *Journal of the American Society for Information Science* 42 (1991), 7-15.

[18] A. Pirkola and K. Järvelin, Employing the resolution power of search keys, *Journal of the American Society for Information Science and Technology*, 52(2001), 575−583.

[19] A. Loponen, and K. Järvelin, A dictionary- and corpus-independent statistical lemmatizer for information retrieval in low resource languages. In: Agosti, M, Ferro, N., Peters, C., de Rijke, M., Smeaton, A. (eds.) CLEF'10 Proceedings of the 2010 international conference on Multilingual and multimodal information access evaluation: cross-language evaluation forum. LNCS vol. 6360, (2010), 3–14. Springer, Heidelberg.

[20] K.W. Church, The DDI approach to morphology. In: Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday. Editors: A. Arppe, L. Carlson, K. Lindén, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund and A. Yli-Jyrä, (2005), 25-34. http://cslipublications.stanford.edu/koskenniemi-festschrift/kk-festschrift-all-2005.pdf.

[21] P. Grünwald, *The Minimum Description Length Principle*. MIT Press, 2007.