Don’t Have a Stemmer? Be Un+concern+ed

Paul McNamee† (JHU), Charles Nicholas‡ (UMBC), James Mayfield† (JHU)

†Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore MD USA

‡Department of Computer Science and Electrical Engineering, UMBC, Baltimore MD USA

Introduction

The use of stemming to address morphological variation is pervasive.

Advantages of Stemming	Disadvantages of Stemming
Increased recall	Mistakes (less precision)
Reduction in lexicon size	Extra time & effort
	Not universally available

In a multilingual context where there are documents in multiple languages stemming is harder to implement. Rule-based stemming tools are popular and well-studied [3,4,5], but not available for many languages. Even when stemmers are available, they are unlikely to have a software implementation in a common software language or API. Therefore it is worth considering statistical stemmers, which can learn to normalize surface forms based on a sample of text alone; however, in some languages, stemming is of lesser importance because only a small number of inflectional forms is used.

This paper compares three methods of word tokenization for information retrieval: (1) rule-based stemming using the *Snowball* stemmer [8]; (2) word segments produced by an unsupervised morphological analysis tool, *Morfessor* [1]; and, (3) fixed-length character n-grams with n=4 and n=5 [7]. Our goal is to explore whether: “alternatives to rule-based stemming successfully improve IR effectiveness using unnormalized word forms?” A recent evaluation at the Morphology Challenge workshop in 2007 compared a variety of methods for unsupervised morphological analysis [6], and motivated this study.

Tokenization

Snowball applies a cascade of rules to normalize word forms. For example, rules like

`ly$ → li` and `bli$ →ble` and `le$ → l`

could map “possible” and “possibly” to “possibl”. Snowball is available in a variety of programming languages and may be obtained from http://snowball.tartarus.org/

Morfessor takes as input a word list and attempts to find an optimal model based on the minimum description length (MDL) principle, which balances the length of the model codebook and the fit of the model on the observed data. Morfessor produces a segmentation for each word, for example, ‘affectionate’ is split into three pieces: affect+ion+ate. No letter substitutions occur, so the verbs “fly” and “fli+es” will not match. The algorithm is completely language neutral and is suited for concatenative morphology. During indexing every segment was added to the inverted file.

Character n-grams transform input words into a set of substrings that each share n-1 characters with the previous n-gram. For example, with n=5, “isle of man” would be represented with {_isle, isle_, sle_o, le_of, e_of_, _of_m, of_ma, f_man, _man_}. N-grams achieve morphologic regularization indirectly due to the fact that subsequences that touch on word roots will match. For example, “juggling” and “juggler” will share the 5-grams _jugg and juggl. While n-gram’s redundancy enables useful matches, other matches are less valuable, for example, every word ending in ‘tion’ will share 5-gram tion_ with all of the others; however, in practice these ‘morphological false alarms’ are almost completely discounted.

Word	Snowball	Morfessor	5-grams
authored	author	author+ed	_auth, autho, uthor, thore, hored, ored_
authorized	author	author+ized	_auth, autho, uthor, thori, horiz, orize, rized, ized_
authorship	authorship	author+ship	_auth, autho, uthor, thors, horsh, orshi, rship, ship_
reauthorization	reauthor	re+author+ization	_reau, reaut, eauth, autho, uthor, thori, horiz, oriza, rizat, izati, zatio, ation, tion_
afoot	afoot	a+foot	_afoo, afoot, foot_
footballs	footbal	football+s	_foot, footb, ootba, otbal, tbaall, balls, alls_
footloose	footloos	foot+loose	_foot, footl, ootlo, otloo, tloos, loose, oose_
footprint	footprint	foot+print	_foot, footp, ootpr, otpri, tprin, print, rint_
feet	feet	feet	_feet, feet_
juggle	juggl	juggle	_jugg, juggl, uggle, ggle_
juggled	juggl	juggle+d	_jugg, juggl, uggle, ggled, gled_
jugglers	juggler	juggle+r+s	_jugg, juggl, uggle, ggler, glers, lers_

Experimental Setup

Data

We examined 13 languages using data from the Cross-Language Evaluation Forum (CLEF) ad hoc test sets [2]. The corpora consist of newspaper documents between 2002 and 2007. We used up to two year’s worth of documents and queries per language.

Retrieval

The JHU HAIRCUT system was used with a language model similarity metric. Document term frequencies were smoothed using the corpus by linear interpolation and a smoothing constant of 0.5. Automated relevance feedback was not employed.

Evaluation

We choose mean average precision (MAP) as the evaluation measure and conducted tests of statistical significance with the Wilcoxon test.

Results

In Table 1 we report mean average precision for words, Morfessor segments, Snowball stems, and character 4-grams. Performance of 5-grams (not shown) is quite similar to 4-grams.

Language	# Docs	Words	Morfessor	Snowball	4-grams
Bulgarian	85,427	0.2195	0.2786 (+26.9%)		0.3163 (+44.1%)
Czech	81,735	0.2270	0.3215 (+41.6%)		0.3294 (+45.1%)
Dutch	190,605	0.4162	0.4274 (+2.7%)	0.4273 (+2.7%)	0.4378 (+4.9%)
English	166,754	0.4829	0.4265 (-11.7%)	0.5008 (+3.7%)	0.4411 (-8.7%)
Finnish	55,344	0.3191	0.3846 (+20.5%)	0.4173 (+30.7%)	0.4827 (+51.3%)
French	129,804	0.4267	0.4231 (-0.84%)	0.4558 (+6.8)%)	0.4442 (+4.1%)
German	294,805	0.3489	0.4122 (+18.1%)	0.3842 (+10.1%)	0.4281 (+22.7%)
Hungarian	49,530	0.1979	0.2932 (+48.2%)		0.3549 (+79.3%)
Italian	157,558	0.3950	0.3770 (-4.6%)	0.4350 (+10.1%)	0.3925 (-0.6%)
Portuguese	210,734	0.3232	0.3403 (+5.3%)		0.3316 (+2.6%)
Russian	16,715	0.2671	0.3307 (+23.8%)		0.3406 (+27.5%)
Spanish	454,041	0.4265	0.4230 (-0.82%)	0.4671 (+9.5%)	0.4465 (+4.7%)
Swedish	142,819	0.3387	0.3738 (+10.4%)		0.4236 (+25.1%)
Average		**0.3376**	**0.3701 (+9.6%)**	**0.3614 (+7.0%)**	**0.3976 (+17.7%)**

Table 1. Performance for four tokenization types in 13 languages. Segments achieved more than a 20% improvement in Bulgarian, Finnish, and Russian, and over 40% in Czech and Hungarian. Snowball stems could not be computed in Bugarian, Czech, Hungarian, Portuguese, or Russian. Both segments and stems improve on unnormalized words, but 4-grams do best of all.

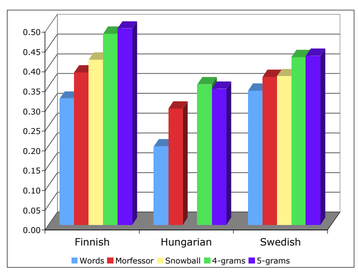

Figure 1. Mean average precision for five tokenization types in three selected languages. The trend is that 4-grams and 5-grams are comparable, Snowball stems (when available) are a bit worse, but better than Morfessor segments which outperform words.

Conclusions

Morfessor segments are effective

Unsupervised segmentation brought a 9.6% improvement in retrieval effectiveness compared to plain words. In languages with high morphological complexity large gains (from 20 to 48%) were observed.

Stemming works better in low complexity languages

When rule-based stemming was available, it outperformed segments.

Character n-grams perform best

Outside the Romance family, where Snowball stems had an advantage, character n-grams exhibited the best performance. 4-grams and 5-grams were about equally effective.

References

[1] M. Creutz and K. Lagus. Unsupervised discovery of morphemes. In ACL-02 Workshop on Morphological and Phonological Learning, pages 21–30, 2002.

[2] G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. CLEF 2007: Ad hoc track overview. In CLEF 2007 Working Notes, 2007.

[3] D. Harman. How effective is stemming? JASIS, 42(1):7–15, 1991.

[4] D. A. Hull. Stemming algorithms: A case study for detailed evaluation. JASIS, 47(1):70–84, 1996.

[5] R. Krovetz. Viewing morphology as an inference process. In ACM SIGIR 1993, pages 191–202, 1993.

[6] M. Kurimo, M. Creutz, and V. Turunen. Overview of Morpho Challenge in CLEF 2007. In Working Notes of the CLEF 2007 Workshop, 2007.

[7] P. McNamee and J. Mayfield. Character n-gram tokenization for european language text retrieval. Information Retrieval, 7(1-2):73–97, 2004.

[8] M. F. Porter. An algorithm for suffix stripping. Program, 14:130–137, 1980.

For further information

Please contact paul.mcnamee@jhu.edu.

UMBC

AN HONORS UNIVERSITY IN MARYLAND

JOHNS HOPKINS

UNIVERSITY