Hungarian Monolingual Retrieval at CLEF 2005

Anna Tordai Maarten de Rijke
Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
atordai,mdr@science.uva.nl

Abstract

We describe our official runs for the ad hoc monolingual task in Hungarian for CLEF 2005. We conducted experiments with four stemmers of varying impact. The experiments indicate that stemmers focusing on noun inflection are as effective as more broadly oriented stemmers, and that extensive stemming is especially beneficial for Hungarian monolingual retrieval.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Stemming, Morphological analysis, Hungarian language

1 Introduction

In our participation in the CLEF ad hoc task this year, we focused exclusively on monolingual retrieval for Hungarian. This is the first year Hungarian is part of CLEF and it is an ideal opportunity to test our work on the effects of stemming in Hungarian. Previous work on languages that are morphologically richer than English, such as Finnish, indicate that there should be benefits from morphological analysis such as stemming, lemmatization, and compound analysis [3, 4, 5]. We have developed a number of suffix-stripping algorithms of varying impact, all focusing on inflectional suffixes. Our goal is to determine the degree of stemming that would prove beneficial for retrieval effectiveness, in terms of both precision and recall. We expect to see improvements for recall for the stemmers, but in addition, we hope that our "light" stemmers keep precision at an acceptable level. The "heavy" stemmer we developed is also expected to improve recall, but it will probably hurt precision.

The paper is organized as follows. Section 2 describes the traits of the Hungarian language that are important from an information retrieval point of view. Section 3 contains a description algorithmic stemmers along with an evaluation. Section 4 describes the retrieval system we used. Section 5 concerns the official and non-official runs, finally followed by a conclusion.

2 Hungarian Morphology

Hungarian is an agglutinative language remotely related to Finnish and Estonian, and a member of the Ob-Ugric languages [7]. The Hungarian language is highly inflectional, rich in compound words, and has an extensive inflectional and derivational morphology. To illustrate this, nouns have 16 to 24 cases depending on the classification system. Additionally, if person, number and possession are added for a single noun there may be as many as 1400 forms [2]. Adjectives also have case, person, number and possession, as well as degree, pushing the number of forms to around 2700. Verbs have fewer forms, with person, number, tense, transitivity adding up to 59. These numbers merely illustrate the inflectional variety of the language. Additionally, there is an extensive system of derivational suffixes, many of them changing the part of speech of a word.

Compound words are frequent in Hungarian, presenting an additional challenge for retrieval. Compound nouns can be formed by two nouns and a participle and a noun. Adjectives can also be formed by the combination of a noun and adjective. Compounding was not addressed at this time.

3 Algorithmic Stemmers

In this section we describe and evaluate the stemmers used in our retrieval experiments.

3.1 Description of the Stemmers

The stemmers were built in the Snowball language [9] and are rule-based stemmers focusing on inflectional suffixes in Hungarian. Using the Szeged Corpus [1], which is a collection of annotated texts ranging from novels, children's essays, legal texts, newspaper articles to computer books, we created a list of the most frequent types of morphosyntactic tags. This helped to determine which suffixes appear most often in the text and guided the construction of the stemmers.

We developed four types of stemmers:

- *Light1* – handling frequent noun cases, plural and frequent owners.
- *Light2* – handling all noun cases, plural and frequent owners.
- *Medium* – handling frequent noun cases, plural, frequent owners and frequent verb tenses.
- *Heavy* – handling most inflectional suffixes.

We will now discuss the stemmers in more detail.

The lightest stemmer, *Light1*, only handles 14 frequent noun cases, plural and the most frequent possessive cases. It is the least invasive stemmer but statistics suggest it might still have a significant impact. Of all the nouns in the Szeged corpus 26% were in uninflected form. The most frequent types of suffixes cover 36% of the nouns. These were the ones targeted by *Light1* with the exception of the single letter suffix 'k' indicating plurality. Even without it, at least half of all nouns should be indexed in their stem form. Since adjectives have the same case, number and possession suffixes as nouns they also become stemmed along with numerals which also share a number of cases with nouns.

The second stemmer, *Light2*, is similar to *Light1* except that it handles 21 noun cases instead of just 14, also removing single letter suffixes such as the accusative 't' and superessive 'n'. The *Light1* and *Light2* stemmers both take word length into account, making sure the remainder is at least a valid vowel-consonant combination.

The third stemmer, *Medium*, removes 12 frequent noun cases, plural, possession and combinations of ownership and plurality. It also handles frequent verb tense-person-number combinations as well as the degree of adjectives. In addition, suffixes forming ordinals and fractions out of numerals were also removed.

	UI	OI	SW	ERRT
Light1	0.75	0.0000028	0.0000037	0.81
Light2	0.59	0.0000053	0.0000089	0.66
Medium	0.64	0.0000081	0.0000127	0.73
Heavy	0.53	0.0000134	0.000025	0.65

Table 1: Performance of the stemmers on the word-groups

The last stemmer, *Heavy*, is the most aggressive, removing 21 noun cases, handling plurality and possession. For verbs it handles infinitive, indicative, conditional and subjunctive moods.

Unfortunately, there are a number of difficulties for the stemmers such as overstemming and homonymy. As an example of overstemming, the word *nemzet*, meaning 'nation', is already in stemmed form but the heavy stemmer removes the 'et' suffix since it is a valid accusative ending leaving the invalid *nemz* as the stem. This problem could be alleviated by expanding the stemmer with the use of an exceptions list containing certain frequent words.

To illustrate the problem of homonymy, consider the word *nevet*, which either means the verb 'to laugh' or the noun 'name' in accusative form. For the latter the stemmer ought to remove the 'et' ending and swap 'e' for 'é' to produce *név*; for the former it must leave the word untouched. What complicates the decision whether to remove this suffix, is that accusative is the most frequent case in the Szeged corpus after the nominative case. At present the stemmers that remove the accusative case overstem the verb form. It would be interesting to see if a lemmatizer would have an edge over an algorithmic stemmer when it comes to these problems.

3.2 Evaluating the Stemming Algorithms

The stemmers were evaluated both intrinsically and extrinsically. For the intrinsic evaluation, we used Paice's method based on error counting [8]. According to this method, two values determine the quality of a stemmer: *understemming* and *overstemming*. In order to determine these values, a list of words is separated into conceptual groups formed by semantically and morphologically related words. This is the target, and an ideal stemmer should conflate words to these conceptual groups.

The stemmers were used to stem the word list, and following the Paice method their correspondence to the conceptual groups was measured. This resulted in an understemming (UI) and overstemming measure (OI). By dividing the overstemming index by the understemming measure we get the stemming weight (SW) which is a measure of the strength of the stemmer.

Paice also offers a different way to combine the two measures (UI and OI) to determine the general relative accuracy of the stemmers. This measure, called *error rate relative to truncation*, or ERRT, is useful for deciding on the best overall stemmer in cases where one stemmer is better in terms of understemming but worse in terms of overstemming. To calculate the ERRT we created a baseline using length truncation by reducing the words in the world list to their n first letters where n was 9, 10, 11 and 12. The overstemming and understemming measure of these truncated lists defines the truncation line. The values of any reasonable stemmers are found between this line and the origin. Figure 1 shows the UI and OI values for each stemmer with the truncation line. Generally, the further the stemmer is from this line, the better it performs on the word lists. By drawing a line that passes through the origin, the datapoint identified by the pair (UI,OI) consisting of the stemmer's understemming and overstemming index, respectively, and that intersects the truncation line, we obtain the distances necessary to calculate the ERRT value of each stemmer. These are the distance from the origin to the stemmer's (UI,OI) divided by the distance from the origin to the intersection (with the truncation line). Low overstemming and understemming indexes are the desired feature in a stemmer. Stemmers that are closer to the origin have lower UI and OI values which means the distance is also shorter. The 'best' stemmer would also have the lowest ERRT value compared to the rest.

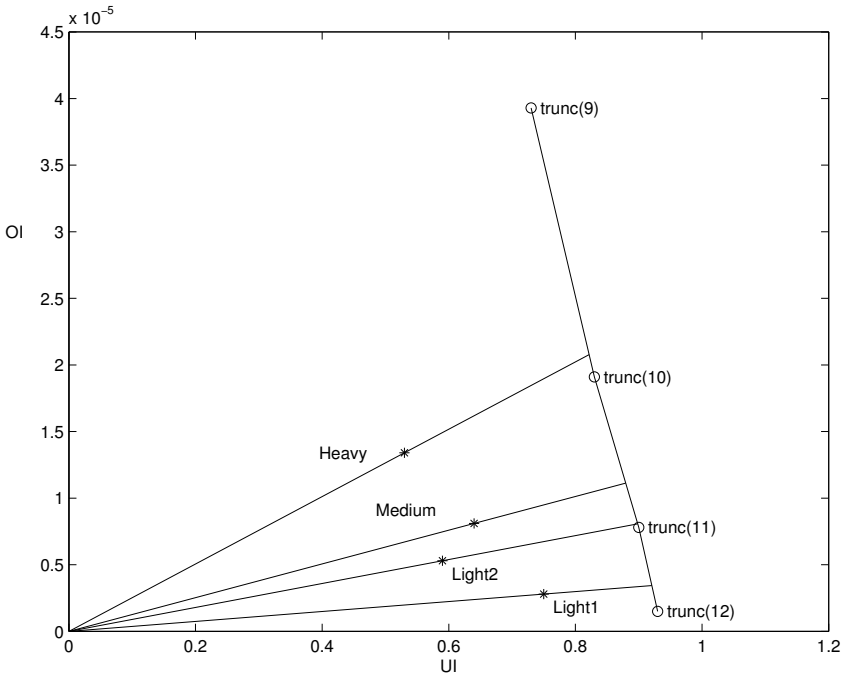

Figure 1: UI × OI plot with the ERRT distances.

Table 1 contains the UI, OI, SW and ERRT values for each of the four stemmers used. As expected, *Light1*, being the lightest stemmer, has the highest understemming index, while *Heavy* has the lowest value. The high value for understemming for *Light1* indicates that it leaves many words unstemmed or just understemmed. The reverse is true for the overstemming index. The *Medium* stemmer has a lower understemming and higher overstemming index than *Light2* which, at first sight, seems surprising. However, 54% of the words in the list are nouns, and since *Light2* removes all noun cases just like the *Heavy* stemmer, but unlike *Light1* and *Medium*, these scores make sense. The *Medium* stemmer focuses on some frequent noun cases and verbs. Verbs only form 23% of the word list so the reason for the somewhat unexpected values is simply due to the fact that the *Medium* stemmer stems fewer words than *Light2*. Overall, when it comes to stemming a word list, a stemmer handling all noun cases yields better results than one restricted to the most frequent noun cases and verb tenses. We suspect that this will apply to a lesser extent for retrieval as words are unique in the word list unlike in a normal corpus.

The high ERRT value of *Light1* indicates that although it has very low overstemming it leaves too many words understemmed making it too light. The same is true for the *Medium* stemmer, which loses out because it focuses on verbs even though there are fewer verbs than nouns in the word list. In this sense *Light2* and *Heavy* come out as winners having the lowest ERRT values. What would this mean when used in an information retrieval setting? An analysis of English topics used in CLEF 2004 showed that after stopping over 65% of the words were nouns, only 10% verbs and 12% adjectives. A post submission analysis confirmed these findings for the 2005 Hungarian topics, with 60% of nouns, 23% adjectives and 17% verbs after stopping. Thus, even if a stemmer only concentrates on stemming nouns it should still have an impact on recall if not precision. Based on the ERRT values we expect the runs with *Light2* and *Heavy* stemmers to yield a better recall than the other two stemmers and the baseline (no stemming at all). At the same time, precision will probably be negatively affected by the *Heavy* stemmer. These results suggest that the run with *Light2* should have the highest recall and precision values since it has a low understemming ratio and should still stem a large percentage of words. Let's see.

4 Retrieval Setup

Now that we have described the stemmers that we have developed, we turn to our retrieval experiments and submissions. First, we used Lucene (off-the-shelf) for indexing and retrieval with a standard vector space model [6].

In addition, we used a stopword list which was created using the Szeged Corpus [1]. We created a list from the 300 most frequent words in the corpus. Numbers and homonyms were removed for the list and it was expanded with pronouns. The result was a list of 188 words.[1] Both the index and queries were stopped. Diacritics were left untouched.

The Hungarian document collection for CLEF 2005 consists of a collection of the newspaper *Magyar Hírlap* from 2002. The document collection was encoded in UTF-8. As the Snowball stemmers were created for ISO Latin encoding the entire collection was converted into ISO Latin 1 encoding without any loss of textual data. For each document the title, lead and description fields were allowed to be used for retrieval; they were all indexed.

There were 50 queries and we used both the title (T) and description (D) fields for retrieval.

5 CLEF 2005 Experiments

In this section describe the results of our experiments.

5.1 Runs

We submitted four official runs for the monolingual Hungarian ad-hoc task, one for each of the four stemmers we developed:

- Light1 (run id: UAmsMoHu4AnV)
- Light2 (run id: UAmsMoHu3AnL)
- Medium (run id: UAmsMoHu2AnG)
- Heavy (run id: UAmsMoHu1AnH)

We conducted several post-submission experiments once the assessments and the results of the submitted runs had been made available:

- Base
- Base + stop
- 6-grams

Some of these additional experiments serve as baseline runs to assess the overall impact of the Snowball stemmers and of stopping. In addition we ran experiments using character n-grams. The run with 6-grams was not stopped; for this run the corpus was indexed with the original word and its 6-grams. The 6-gram performed better than 7-grams and 8-grams, which are not discussed in here.

5.2 Retrieval Results

The Mean Average Precision (MAP) scores in Table 2 show that of the stemmed runs the *Heavy* stemmer performed best, closely followed by *Light2*, while *Medium* and *Light1* perform worse. This confirms the results of the ERRT values in Section 4 and suggests that extensively stemming nouns yields good results and even more extensive stemming improves precision. It is worth noting that the MAP score of the *6-Gram* run is only slightly below the score of *Light2*. In fact,

[1] The stopword list is available at http://ilps.science.uva.nl/Resources/.

	MAP	R-prec	Relevant Docs Retrieved
Light1	0.2150	0.2416	700
Light2	0.2799	**0.2905**	734
Medium	0.2330	0.2556	717
Heavy	**0.2819**	0.2839	740
Base	0.1831	0.2096	591
Base + stop	0.1836	0.2014	607
6-Gram	0.2787	0.2903	**747**

Table 2: Overview of MAP scores and R-precision scores for the official and unofficial runs. Best scores are in bold face.

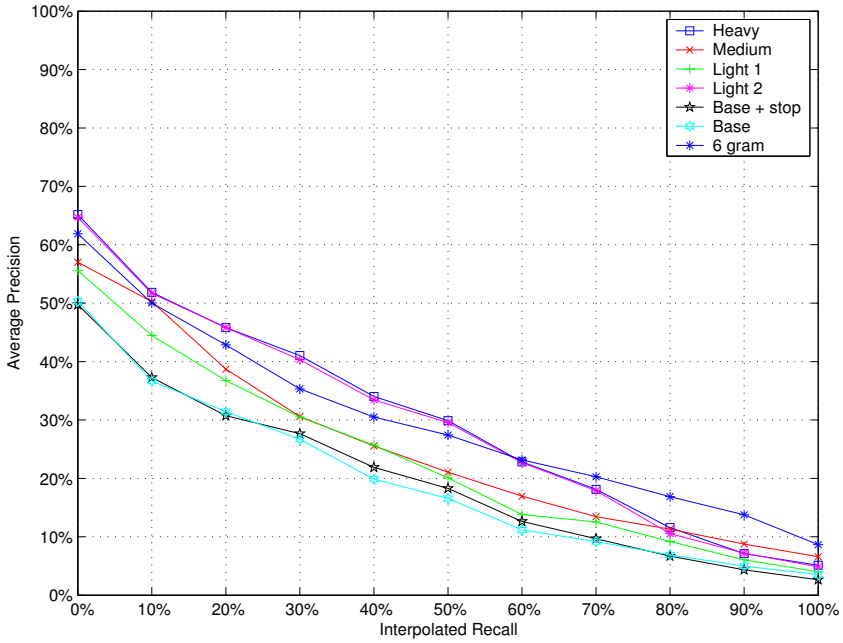

Figure 2: Interpolated Recall vs. Average Precision for official and unofficial runs.

we see that using 6-grams on this corpus is almost as good as using stemming. When looking at the Interpolated Recall vs Average Precision (Figure 2) we see that the *Light2* run has almost exactly the same values as the *Heavy* run. When it comes to R-precision scores *Light2* has the best scores beating the *Heavy* run.

Interestingly, the *6-Gram* run retrieves the largest number of relevant documents but does not rank them sufficiently high, as the MAP scores are not the best ones. Judging from the combination of number of retrieved relevant documents, R-precision and MAP scores, *Light2* ranks the relevant documents the highest even though it does not retrieve as many as the *6-gram* or *Heavy* runs.

5.3 Discussion

The retrieval performance of the stemmed runs follows our expectations from Section 4. Of the four stemmers, the *Heavy* and *Light2* stemmers performed best, followed by *Medium* and *Light*. The *Light2* stemmer shows the importance of stemming nouns when it comes to retrieval in Hungarian. The *Heavy* stemmer indicates that even extensive stemming does not lower precision like it is known to do for English retrieval.

We will now look at the performance of the stemmed runs on certain topics. As in general, the *Heavy* and *Light2* stemmers performed much better per topic than the other two stemmers did. We will compare the stronger group to the weaker group in order to determine why there was such a difference in performance as well as why the stemmers performed below the median.

Topic 289 is an example where all of the runs were below the median. In this document the task is to retrieve documents about the Falkland islands.

<num> C289 </num>
<HU-title> Falkland-szigetek </HU-title>
<HU-desc> Keressünk a Falkland-szigetekről szóló cikkeket.</HU-desc>

Some of the relevant documents were not retrieved. One of the reasons for this was that the term *Falkland-szigetek* was indexed as a single word. Hyphenated words are frequent in Hungarian for dates, acronyms and when foreign words or brand names become inflected (e.g., *for NATO* becomes *NATO-nak*). One of the relevant documents contained the words *Falkland-szigetek* ("Falkland islands") in separate form while another contained the term *Falkland-háború* ("Falkland wars") and these were not retrieved.

Another document was not found because it contained the adjective form *Falkland-szigeti* meaning "from the Falkland islands." This is a derivative suffix and, as mentioned earlier, derivative suffixes are not removed by our stemmers. However, this type of suffix is so frequent it will be removed in future versions of the stemmer.

One of the topics where the *Heavy* and *Light2* runs did much better (noticeably higher than the median) than the weaker ones was Topic 259. Documents relevant for this topic contain information about movies that have been awarded the Golden Bear at the Berlin film festival.

<num> C259 </num>
<HU-title> Aranymedve </HU-title>
<HU-desc> Mely filmek kaptak Aranymedve díjat a berlini filmfesztiválon? </HU-desc>

While all four relevant documents were retrieved in each run it was their ranking that resulted in different precision scores. The *Heavy* and *Light2* stemmers correctly stemmed one form of the word "Aranymedve" in one of the relevant documents as well as the word "filmfesztivál". This boosted the ranking of these documents. The word "film" was correctly stemmed by all stemmers in all the documents but as this word appears frequently in irrelevant document as well it resulted in lower rankings for the *Medium* and *Light1* runs. It is worth noting that the word *filmfesztiválon* is a compound of *film* and *fesztiválon*. Although the word *film* was in the query, we believe that decompounding the word *fesztiválon* would have helped to would have further boosted the ranking.

Now that we know the form of the topics we will also be able to adjust the stopword list so as to include the words *keressünk* (search) and *cikkeket* (articles).

6 Conclusion

The experiments on which we report in this paper confirm that stemming in Hungarian greatly improves retrieval effectiveness. They show that a stemmer focusing merely on the inflection of nouns works almost as well as a more broadly oriented stemmer. Merely stemming frequent noun and verb inflections however yields worse results than using 6-grams. Our results are sobering as a 6-grammed run performed almost as well as the best performing stemmed run.

Our stemmers themselves can be improved upon and hyphenated words will have to be addressed differently in the future. A detailed error analysis has shown that decompounding will probably boost rankings and help retrieve additional documents. Analyzing the impact of decompounding on Hungarian monolingual retrieval is left as future work, though.

7 Acknowledgements

Maarten de Rijke was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 612-13-001, 612.000.106, 612.000.207, 612.066.302, 612.069.006.

References

[1] Szeged Corpus. A morpho-syntactically annotated and POS tagged Hungarian corpus, 2005.

[2] T. Erjavec and M. Monachini. Specifications and notation for lexicon encoding. Technical report, COP Project 106 MULTEXT - East, December 17, 1997.

[3] S. Fissaha Adafre, W.R. van Hage, J. Kamps, G. Lacerda de Melo, and M. de Rijke. The University of Amsterdam at CLEF 2004. In C. Peters and F. Borri, editors, *Working Notes for the CLEF 2004 Workshop*, pages 91–98, 2004.

[4] V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual document retrieval for European languages. *Information retrieval*, 7:33–52, 2004.

[5] T. Korenius, J. Laurikkala, K. Jarvelin, and M. Juhola. Stemming and lemmatization in the clustering of finnish text documents. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management*, 2005, pages 625–633.

[6] Lucene. The Lucene search engine. URL: http://jakarta.apache.org/lucene/.

[7] B. Megyesi. The Hungarian language. URL: http://www.speech.kth.se/~bea/hungarian.pdf.

[8] C.D. Paice. Method for evaluation of stemming algorithms based on error counting. *Journal of The American Society for Information Science*, 47(8):632–649, 1996.

[9] Snowball. The Snowball string processing language. URL: http://snowball.tartarus.org/, 2005.