# Bulgarian, Hungarian and Czech Stemming Using YASS

Prasenjit Majumder, Mandar Mitra, and Dipasree Pal

CVPR Unit, Indian Statistical Institute, Kolkata
{prasenjit_t,mandar,dipasree_t}@isical.ac.in

**Abstract.** This is the second year in a row we are participating in CLEF. Our aim is to test the performance of a statistical stemmer on various languages. For CLEF 2006, we tried the stemmer on French [1]; while for CLEF 2007, we did experiments for the Hungarian, Bulgarian and Czech monolingual tasks. We find that, for all languages, YASS produces significant improvements over the baseline (unstemmed) runs. The performance of YASS is also found to be comparable to that of other available stemmers for all the three east European Languages.

## 1 Introduction

Stemming is arguably a recall enhancing device in text retrieval. Most commonly used stemmers are rule-based and therefore language specific. Such stemmers are unlikely to be available for resource poor languages. In earlier work, therefore, we proposed YASS [2], a statistical stemmer. As YASS does not assume any language specific information, we expect the approach to work for multiple languages. The motivation behind our experiments at CLEF 2006 last year was to test this hypothesis. Since our hypothesis was supported by last year's experiments, this year, for CLEF 2007, we planned on monolingual retrieval for more languages which we know nothing about.

The main stumbling block in our experiments was the encoding issue. We modified our systems to work with UTF-8 data. During the official submission, we could not complete the Bulgarian runs and submitted only six official runs for Hungarian and Czech. After the relevance judgements were released, we tuned the statistical stemmer for each of the three languages.

Three retrieval models were used in our study, viz. BM25, DFR-In_expC2, and TF.IDF (Lnu.ltn). Our experiments were conducted using the SMART[3] system for the tf.idf model, and the Terrier[4] system for the rest of the models.

We give a brief overview of YASS in the next section. Section 3 presents and analyses the results of all the runs (both official and unofficial) for the three languages. Section 4 concludes the paper.

## 2 YASS

YASS (Yet Another Suffix Stripper) [2] is a statistical stemmer that is based on a string distance measure. Using this measure, YASS clusters a lexicon created

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 49–56, 2008.

from a text corpus. Each cluster is expected to contain all the morphological variations of a root word. The clustering method (agglomerative hierarchical clustering) requires a threshold value (refered to as $\theta$ henceforth) as a parameter. If training data is available, this parameter may be tuned to improve the performance of the stemmer. The following subsections will describe the string distance used and the training procedure for threshold selection.

### 2.1 String Distance Measures

Distance functions map a pair of strings $s$ and $t$ to a real number $r$, where a smaller value of $r$ indicates greater similarity between $s$ and $t$. In the context of stemming, an appropriate distance measure would be one that assigns a low distance value to a pair of strings when they are morphologically similar, and assigns a high distance value to morphologically unrelated words. The languages that we have been experimenting with are primarily suffixing in nature, i.e. words are usually inflected by the addition of suffixes, and possible modifications to the tail-end of the word. Thus, for these languages, two strings are likely to be morphologically related if they share a long matching prefix. Based on this intuition, we define a string distance measure $D$ which rewards long matching prefixes, and penalizes an early mismatch.

Given two strings $X = x_0x_1 \ldots x_n$ and $Y = y_0y_1 \ldots y_{n'}$, we first define a Boolean function $p_i$ (for penalty) as follows:

$$p_i = \begin{cases} 0 & \text{if } x_i = y_i \quad 0 \leq i \leq \min(n, n') \\ 1 & \text{otherwise} \end{cases}$$

Thus, $p_i$ is 1 if there is a mismatch in the $i$-th position of $X$ and $Y$. If $X$ and $Y$ are of unequal length, we pad the shorter string with null characters to make the string lengths equal.

Let the length of the strings be $n + 1$, and let $m$ denote the position of the first mismatch between $X$ and $Y$ (i.e. $x_0 = y_0,\ x_1 = y_1,\ \ldots,\ x_{m-1} = y_{m-1},$ but $x_m \neq y_m$). We now define $D$ as follows:

$$D(X, Y) = \frac{n - m + 1}{m} \times \sum_{i=m}^{n} \frac{1}{2^{i-m}} \quad \text{if } m > 0, \quad \infty \text{ otherwise} \tag{1}$$

Note that $D$ does not consider any match once the first mismatch occurs. The actual distance is obtained by multiplying the total penalty by a factor which is intended to reward a long matching prefix, and penalize significant mismatches. For example, for the pair $\langle$*astronomer*, *astronomically*$\rangle$, $m = 8, n = 13$. Thus, $D = \frac{6}{8} \times (\frac{1}{2^0} + \ldots + \frac{1}{2^{13-8}}) = 1.4766$.

### 2.2 Lexicon Clustering

Using the distance function defined above, we can cluster all the words in a document collection into groups. Each group, consisting of "similar" strings, is

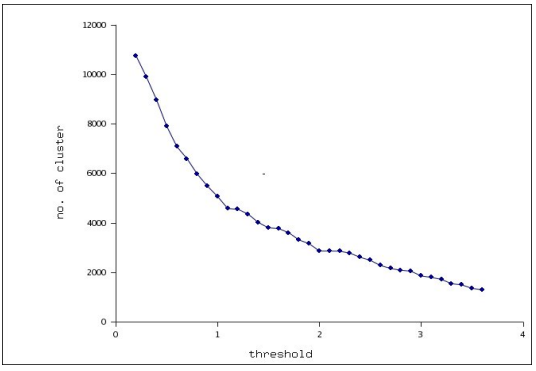


**Fig. 1.** Number of clusters at various thresholds for Hungarian

expected to represent an equivalence class consisting of morphological variants of a single root word. The words within a cluster can be stemmed to the 'central' word in that cluster. Since the number of natural clusters are unknown apriori, partitive clustering algorithms like $k$-means are not suitable for our task. Also, the clusters are likely to be of non-convex nature. Graph-theoretic clustering algorithms appear to be the natural choice in this situation because of their ability to detect natural and non-convex clusters in the data.

Three variants of graph theoretic clustering are popular in the literature, namely, *single-linkage*, *average-linkage*, and *complete-linkage*. We choose the compete-linkage algorithm for our experiments.

### 2.3 Training

We have mentioned earlier that YASS needs no linguistic input as it is a statistical stemmer. However, before running YASS on a new language, we need to train it for getting a suitable clustering threshold. As training data was not available for Bulgarian and Czech before the official submission, we set the threshold value to 1.5 based on our earlier experience with English and French. For Hungarian, we used the CLEF2006 data for training. A lexicon was extracted from the corpus and clustered using various thresholds, resulting in a set of stemmers. A suitable threshold was chosen based on the performance of these stemmers. After the relevance judgements were released for all the languages, we tuned the threshold of YASS for Bulgarian and Czech as well, using this year's data.

**Hungarian.** The same Hungarian corpus is used for the 2005, 2006, and 2007 tasks. The lexicon extracted from the corpus has 536,678 surface words. The lexicon was clustered using various threshold settings, and the number of clusters versus threshold curve is shown in Figure 1. The step-like regions around 0.8, 1.1, 1.5, 2.0 suggest that the number of clusters is stable around these threshold values. These values may thus be chosen as candidate thresholds for clustering. After clustering the lexicon using these four threshold values, the lexicon size

**Table 1.** Threshold vs. MAP for Hungarian

| Threshold | Mean AvgP (MAP) |
|---|---|
| 0.8 | 0.2692 |
| 1.1 | 0.2835 |
| 1.5 | 0.2777 |
| 2 | 0.2735 |

**Table 2.** Hungarian runs for training on CLEF 2005 dataset

| YASS | | | |
|---|---|---|---|
| **Run Name** | **MAP** | **R-prec** | **% Rel_Ret** |
| noStem(T+D+N) | 0.2472 | 0.2531 | 72.8434 |
| $\theta$ = 0.8(T+D+N) | 0.3211 | 0.3231 | 83.8125 |
| $\theta$ =1.1(T+D+N) | 0.3246 | 0.3247 | 86.0489 |
| $\theta$ =1.5(T+D+N) | 0.3179 | 0.3190 | 86.6879 |
| $\theta$ =2.0(T+D+N) | 0.3005 | 0.3068 | 83.8125 |
| noStem(T+D) | 0.2170 | 0.2285 | 69.1160 |
| $\theta$ =0.8(T+D) | 0.3121 | 0.3162 | 81.0436 |
| $\theta$ =1.1(T+D) | 0.3241 | 0.3270 | 84.2385 |
| $\theta$ =1.5(T+D) | 0.3268 | 0.3309 | 85.6230 |
| $\theta$ =2.0(T+D) | 0.3048 | 0.3074 | 84.1320 |
| TORDAI stemmer | | | |
| **Run Name** | **MAP** | **R-prec** | **% Relevant Docs Retrieved** |
| Heavy minus hyphen | 0.3099 | 0.3048 | 83.1 |
| 4-Gram | 0.3303 | 0.338 | 83.6 |
| 5-Gram | 0.3002 | 0.3057 | 82.4 |

gets reduced to 225489, 169619, 130278, and 76782 classes respectively. The stemmers thus prepared are used in four different official runs.

The official topics for the Hungarian monolingual run at CLEF-2006 were topic numbers 301 to 325 and 351 to 375. Table 1 suggests that the performance of YASS does not change much as the threshold varies between 1.1 and 2.

We also tested these stemmers on CLEF queries 251 to 300. These queries were used in the CLEF 2005 monolingual Hungarian task. Table 2 gives the results of these runs, as well as the best results reported by Tordai et al. [5] for the same task at CLEF 2005. This table also suggests that setting $\theta = 1.1$ or 1.5 would be appropriate for Hungarian.

## 3 Experiments

This section describes all the runs we performed for all the three languages. Besides the six official runs, we performed several other experiments using the three east European languages, to better understand the performance of YASS.

**Table 3.** Official results on 2007 CLEF data

| Hungarian Runs Submitted (*nnn.ltn*) | | | |
|---|---|---|---|
| **Run Name** | **MAP** | **R-prec** | **% Rel_Ret** |
| ISI.YASSHUN | 0.1712 | 0.1974 | 72.22 |
| ISI.YASSTDHUN | 0.1695 | 0.1943 | 72.88 |
| ISI.ISIDWLDHSTEMGZ | 0.1605 | 0.1858 | 66.84 |
| Runs Submitted (*Lnu.ltn*) | | | |
| ISI.CZTD [YASS] (T+D) | 0.3224 | 0.3102 | 87.13 |
| ISI.ISICL [dnlded] (T+D+N) | 0.3362 | 0.3326 | 89.37 |
| ISI.ISICZNS [nostem] (T+D+N) | 0.2473 | 0.2540 | 76.64 |

**Table 4.** Word stems generated by YASS

| Hungarian | | Czech | |
|---|---|---|---|
| politikusokról, politikai | politi | Kosteličových, Kosteličovi | Kostelič |
| atomhulladékot | atomhulladék | prezidenští, prezidenta, prezidentského | preziden |
| megszűnése | megsz | kandidáti, kandidáta | kandidát |
| elnökjelöltek, elnökjelölt | elnökjelöl | vesmírní, vesmírných, vesmíru | vesmír |
| királynő, királyságbeli | kir / király | turistech, turisté | turist |

### 3.1 Official Runs

In the first Hungarian run *ISI.YASSTDHUN*, we indexed only the `<title>` and `<desc>` fields of the queries. For the second run, *ISI.YASSHUN* we indexed the `<title>`, `<desc>`, and `<narr>` fields of the queries. In both cases the clustering threshold was set to 1.5. For the third run, *ISI.ISIDWLDHSTEMGZ*, we made use of a Hungarian stemmer available from the web[1].

The Czech runs are analogous: the first run uses only the `<title>` and `<desc>` fields; the second and third runs use the complete query. The second run makes use of an existing stemmer[2] instead of YASS. The final run was a baseline run where no stemming was used.

Table 3 shows the results of our official runs. These results confirm our hypothesis that YASS will work for a variety of languages, provided the languages are primarily suffixing in nature. Table 4 provides some examples of words and their roots obtained using YASS. These words were selected from queries on which the stemmed run significantly outperformed the unstemmed run.

### 3.2 Other Runs

Other groups that have reported results for these three east European languages in this volume include [6], [7], and [8]. We were particularly interested in the

[1] http://snowball.tartarus.org/algorithms/hungarian/stemmer.html
[2] http://members.unine.ch/jacques.savoy/clef/index.html

**Table 5.** Performance of YASS for various models and parameter settings

| Hungarian | | | | |
|---|---|---|---|---|
| Models | topics | 1.5 | 2.0 | no-stem |
| DFR | TD | 0.3358 | 0.3535 | 0.2461 |
| | TDN | 0.3728 | 0.3902 | 0.2813 |
| OKAPI | TD | 0.2920 | 0.3138 | 0.1992 |
| | TDN | 0.3274 | 0.3445 | 0.2285 |
| TFIDF | TDN | 0.3600 | 0.3638 | 0.2647 |
| Czech | | | | |
| Models | topics | 1.5 | 2.0 | no-stem |
| DFR | TD | 0.3337 | 0.3483 | 0.2320 |
| | TDN | 0.3574 | 0.3674 | 0.2525 |
| OKAPI | TD | 0.3199 | 0.3306 | 0.2162 |
| | TDN | 0.3332 | 0.3464 | 0.2454 |
| TFIDF | TDN | 0.3390 | 0.3381 | 0.2473 |
| Bulgarian | | | | |
| Models | topics | 1.5 | 2.0 | no-stem |
| DFR | TD | 0.3533 | 0.3526 | 0.2586 |
| | TDN | 0.3626 | 0.3649 | 0.2862 |
| OKAPI | TD | 0.3289 | 0.3330 | 0.2346 |
| | TDN | 0.3439 | 0.3465 | 0.2594 |

results reported by Dolamic and Savoy [6] for two reasons. First, their work motivated us to explore retrieval models besides the traditional tf.idf method implemented in the SMART system. Secondly, they present results obtained using linguistically-based stemming / decompounding algorithms. It would be interesting to compare the performance of these methods with that of a purely statistical method such as YASS.

Accordingly, after the relevance judgments for the data sets were distributed, we performed some additional experiments for the three languages. The primary aim of these experiments was two-fold: (i) To use YASS with alternative retrieval approaches, specifically the BM25 weighting method, and the Divergence from Randomness (DFR) model. (ii) To compare YASS with the stemming / decompounding methods described by Dolamic and Savoy.

For these experiments, we used the BM25 scheme and a variant of the Divergence from Randomness model (DFR-In_expC2) as implemented in the Terrier-2.0 system. The $c$ parameter of DFR-In_expC2 was set to the Terrier default value 1.0 for most runs (see Table 5); however, when comparing results with those reported by Dolamic and Savoy, we used $c = 1.5$ as this was the $c$ value used in their work (see Table 6).

Besides exploring alternative retrieval strategies, we also tried a range of clustering thresholds. The results for $\theta = 1.5$ and $\theta = 2.0$ are reported in Table 5. These experiments suggest that 2.0 is a good choice of the parameter $\theta$ in YASS for the three east European languages, irrespective of retrieval models.

### 3.3 Comparing and Analysis of Results

With the clustering threshold $\theta$ set to 2.0, we compared YASS with the stemmers described in [6]. We chose the 12 best runs from that paper for comparison (the 4-gram based runs are not considered, since this approach was less effective than the other approaches). All these runs are based on the DFR model, since it yielded the best performance reported in [6]. As mentioned above, we use $c = 1.5$ for these runs (as suggested in [6]); however, the mean document length (mean dl) parameter is unchanged from the default setting in Terrier (this parameter was set to 213, 135, 152 for Czech, Bulgarian and Hungarian, resp., in [6]).

Table 6 compares the results obtained using YASS with those reported by Dolamic and Savoy. The performance differences were found to be statistically insignificant (based on a $t$-test) for the four Czech and Bulgarian runs.

Of the best four Hungarian runs reported in [6], two runs (TD, TDN) use a stemmer [9,10], and two runs (TD, TDN) use a de-compounding algorithm [11]. Once again, no significant difference was found between these methods and YASS when only the title and description fields of the query were indexed (runs labeled TD). However, the decompounding run using the full query (TDN) was found to be significantly better than YASS. A more detailed analysis of this difference

**Table 6.** Comparison between YASS and Dolamic et al.

| Bulgarian runs | | | | |
|---|---|---|---|---|
| Model | topics | light/word | deriv./word | YASS |
| DFR | TD | 0.3423 | 0.3606 | 0.3613 |
| | TDN | 0.3696 | 0.3862 | 0.3748 |
| Czech runs | | | | |
| Model | topics | light | derivational | YASS |
| DFR | TD | 0.3437 | 0.3342 | 0.3523 |
| | TDN | 0.3678 | 0.3678 | 0.3702 |
| Hungarian runs | | | | |
| Model | topics | stemmer(word) | de-compound | YASS |
| DFR | TD | 0.3525 | 0.3897 | 0.3588 |
| | TDN | 0.4031 | 0.4271 | 0.3951 |

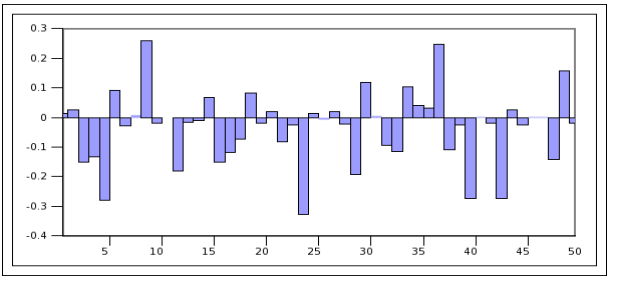


**Fig. 2.** Difference in AvgP for individual queries for YASS and de-compounding

is presented in Figure 2, which shows that, out of 50 queries, YASS performed better in 21 cases, while the decompounding method did better in 29 cases.

## 4 Conclusion

Overall, we found that YASS performs as well as any linguistic stemmers for the three east European languages viz. Hungarian, Bulgarian and Czech. Our explorations of alternative retrieval approaches (besides the traditional tf.idf method) yielded promising results. In future work, we hope to undertake a more complete investigation of YASS within the context of these models.

## References

1. Majumder, P., Mitra, M., Datta, K.: Statistical vs. rule-based stemming for monolingual french retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 107–110. Springer, Heidelberg (2007)
2. Majumder, P., Mitra, M., Parui, S.K., Kole, G., Mitra, P., Datta, K.: Yass: Yet another suffix stripper. ACM Trans. Inf. Syst. 25(4), 18 (2007)
3. Salton, G. (ed.): The SMART Retrieval System—Experiments in Automatic Document Retrieval. Prentice Hall Inc., Englewood Cliffs (1971)
4. Ounis, I., Lioma, C., Macdonald, C., Plachouras, V.: Research directions in terrier. In: Baeza-Yates, R., et al. (eds.) Novatica/UPGRADE Special Issue on Web Information Access (Invited Paper) (2007)
5. Tordai, A., de Rijke, M.: Four Stemmers and a Funeral: Stemming in Hungarian at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 179–186. Springer, Heidelberg (2006)
6. Dolamic, L., Savoy, J.: Stemming approaches for east european languages. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 37–44. Springer, Heidelberg (2008)
7. Ircing, P., Muller, L.: Czech Monolingual Information Retrieval Using Off-The-Shelf Components - the University of West Bohemia at CLEF 2007 Ad-Hoc track. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
8. Ceska, P., Pecina, P.: Charles University at CLEF 2007 Ad-Hoc Track. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
9. Savoy, J., Abdou, S.: Experiments with monolingual, bilingual, and robust retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 137–144. Springer, Heidelberg (2007)
10. Savoy, J.: Searching strategies for the hungarian language. Inf. Process Manage 44(1), 310–324 (2008)
11. Savoy, J.: Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic Models for Effective Monolingual Retrieval. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 322–336. Springer, Heidelberg (2004)