Published in Journal of the American Society for Information Science and Technology, 60, issue 12, 2540-2547, 2009
which should be used for any reference to this work

Indexing and Searching Strategies for the Russian Language

Ljiljana Dolamic and Jacques Savoy
Computer Science Department, University of Neuchatel, Rue Emile Argand 11, 2009 Neuchâtel, Switzerland. {Ljiljana Dolamic, Jacques.Savoy}@unine.ch

This paper describes and evaluates various stemming and indexing strategies for the Russian language. We design and evaluate two stemming approaches, a light and a more aggressive one, and compare these stemmers to the Snowball stemmer, to no stemming, and also to a language-independent approach (*n*-gram). To evaluate the suggested stemming strategies we apply various probabilistic information retrieval (IR) models, including the Okapi, the *Divergence from Randomness* (DFR), a statistical language model (LM), as well as two vector-space approaches, namely, the classical *tf idf* scheme and the *dtu-dtn* model. We find that the vector-space *dtu-dtn* and the DFR models tend to result in better retrieval effectiveness than the Okapi, LM, or *tf idf* models, while only the latter two IR approaches result in statistically significant performance differences. Ignoring stemming generally reduces the MAP by more than 50%, and these differences are always significant. When applying an *n*-gram approach, performance differences are usually lower than an approach involving stemming. Finally, our light stemmer tends to perform best, although performance differences between the light, aggressive and Snowball stemmers are not statistically significant.'

Introduction

Russian belongs to the Indo-European language family and it is the most widely spoken among Slavic languages. Russian is one of three contemporary East Slavic languages (the others being Ukrainian and Belorussian). With 165 million native speakers and 110 million second-language speakers, Russian is among the world's top 10 most spoken languages (Malherbe, 1995), and in Central and Eastern Europe it ranks at the very top. Even though in this region Slavic languages dominate, only a rather small number of document collections are available. For Bulgarian (a South Slavic language), a fairly large collection was created during the 2006 (Peters et al., 2007) and 2007 CLEF campaigns (Peters et al., 2008), while for the Czech language (West Slavic), a test collection was created during the CLEF-2007 campaign (Peters et al., 2008).

In this paper the main objective is to describe the most significant morphological difficulties encountered when applying information retrieval (IR) techniques to the Russian language. Unlike English, where few inflectional suffixes are used to denote number or person variations, Russian makes use of a larger number of them, partly because they are also used to denote grammatical cases (Sproat, 1992). Given the importance of this Slavic language, our goal is to propose, compare, and evaluate various stemming, indexing, and search strategies. Our evaluation will be based on the document collections made available through the 2005 to 2008 CLEF domain-specific tasks. In this case, the main objective is to study information retrieval on domain-specific corpus using both full-text and manual indexing as well the possible usefulness of specialized thesaurus for improving the retrieval effectiveness.

Related Work

In the IR domain (Manning, Raghavan, & Schütze, 2008), it is usually assumed that stemming is an effective means of enhancing retrieval efficiency by conflating several different word variants into a common form. Most stemming approaches achieve this through applying morphological rules for the language involved (for English, see Lovins, 1968; Porter, 1980). In such cases suffix removal is also controlled through the adjunct of quantitative restrictions (e.g., '-ing' would be removed if the resulting stem consisted of more than three letters, as in "running," but not in "king") or qualitative restrictions (e.g., '-ize' would be removed if the resulting stem did not end with 'e' as in "seize"). To improve conflation accuracy, certain ad hoc spelling correction rules are also applied (e.g., "running" becomes "run" and not "runn"), due to certain irregular grammar rules, usually applied to facilitate pronunciation.

Compared to other languages having more complex morphologies (Sproat, 1992), English is considered quite simple, while for other languages such as French simply

applying a dictionary to correct stemming procedures could be more helpful (Savoy, 1993). For those languages having a more complex morphology, deeper analyses could be required (e.g., for Finnish; Korenius, Laurikkala, Järvelin, & Juhola, 2004), and their corresponding lexical stemmers would clearly be more elaborate but they are not always freely available (e.g., Xelda system at Xerox). Not only would their implementation be more labor-intensive and complex, their use would depend on a large lexicon and a complete set of grammar rules for each language involved. This could lead to more processing time and would thus be problematic, especially when document collections are very large and dynamic (e.g., within a commercial search engine on the Web). Additionally, lexical stemmers must be capable of handling unknown words such as geographical, product or proper names, or acronyms (out-of-vocabulary problem). Lexical stemmers thus cannot be viewed as error-free approaches. It must also be recognized that when inspecting language usage and real corpora, the morphological variations observed are less extreme than those involved in grammar. According to Kettunen & Airo (2006), for example, while in theory Finnish nouns have around 2,000 different forms, in current collections most of these forms rarely occur. In fact, 84% to 88% of inflected noun occurrences in Finish are generated by only six out of a possible 14 cases.

While stemming schemes are normally designed to work with general texts, some could also be designed especially for a specific domain (e.g., in medicine) or a given document collection, such as that developed for a corpus-based approach by Xu & Croft (1998). This would more closely reflect language usage (including word frequencies and other co-occurrence statistics) than a set of morphological rules where the frequency of each rule (and therefore its underlying importance) is not precisely known.

Other than English, few stemming procedures have been suggested for European languages (some of them are freely available at snowball.tartarus.org/ or at the Website www.unine.ch/info/clef/). These proposed stemmers usually pertain to the most popular languages, and some of them, like the Finnish language, seem to require a deeper morphological analysis (Korenius et al., 2004) to provide adequate retrieval performances.

Algorithmic stemmers ignore word meanings and also tend to make errors due to over-stemming (e.g., "organization" is reduced to "organ") or to under-stemming (e.g., "European" and "Europe" do not conflate to the same root). Most studies carried out so far involved IR performance evaluations for the English language, while for the less popular languages fewer studies are available. For various European languages, Tomlinson (2004), for example, evaluated the differences between Porter's stemmer (1980) and lexical stemmers (based on a dictionary of the corresponding language). For Finnish and German, the lexical stemmers tend to produce statistically better results, yet for seven other languages the performance differences were insignificant.

Finally, we may also mention the ROMIP evaluation campaigns producing test collections mainly extracted from the Web in the Russian language. However, it was not possible to freely obtain this test collection, and all pertinent information about these corpora, evaluation methodology, and linguistic tools are written in Russian. After analyzing the more recent results, we found that the retrieval performance of the Snowball stemmer tends to reflect the best practice in this field.

Based on these facts, in the rest of this paper we analyze stemmer effectiveness for Russian, and suggest which one would be the most effective. We also address the comparative retrieval effectiveness of an *n*-gram scheme, a language-independent approach, and compare them to a word-based scheme.

Morphology of the Russian Language

When creating stemmers for Russian we started from the same point as for the other languages we have worked with over the past years. We found that the best way to develop effective stemming procedures was to focus mainly on nouns and adjectives (Savoy, 2006), and to avoid verb forms, which are usually too numerous and can lead to a large number of errors.

Russian is a member of the Slavic language family and like many in this family, including Bulgarian, Ukrainian, or Serbian, it is written in the Cyrillic script and uses 33 letters. Other Slavic languages from areas in which Roman Catholicism is the dominant religion, such as Polish and Czech, are written with the Latin alphabet, with various diacritics being added to represent their particular pronunciations.

All Russian nouns have one of three distinct genders (masculine, feminine, or neutral). As in English, all nouns are declined according to number (singular, plural), but some may only have singular or plural forms, as in the English word "scissors." Like most other Slavic languages (except for Bulgarian), all Russian nouns (common or proper nouns) are also declined according to different grammatical cases and we can find six cases in Russian including nominative, genitive, dative, accusative, instrumental, and locative. Each gender-case combination has its own set of characteristic paradigms, including hard-stem types, soft-stem types, and special types. Note, however, that each gender-case combination does not require a distinct suffix. In the first declination, for example, the accusative and the genitive have the same ending, or as shown in Table 1, dative and locative case endings are the same.

TABLE 1. Examples of Russian feminine noun declensions.

		Sister	
Case	Moscow	Singular	Plural
Nominative	Москва	сестра	сёстры
Genitive	Москвы	сестры	сестёр
Dative	Москве	сестре	сёстрам
Accusative	Москву	сестру	сёстры
Instrumental	Москвой	сестрой	сёстрами
Locative	Москве	сестре	сёстрах

Suffixes are not always present and in some cases there are none at all. For example, the feminine noun "book" is written as "книг" in genitive plural and takes the form "книга" in the nominative singular form. The stem is therefore not always the nominative singular (for other examples see Table 1, showing the declension of feminine nouns ending in '-a' in the nominative singular). Usually the stem does not change after adding the required suffix (see Table 1 for a few examples). However, as with other Slavic languages, the presence of a suffix may imply a stem modification, as for example in the elision of the vowel 'o' in the neutral noun "window," which takes the form "окно" in the nominative singular (instead of "оконо" [an incorrect form in Russian] with the final 'o' being a suffix) and "окон" in the genitive plural. This phenomenon is known as the fleeting vowel. Another example we should mention is the vowel 'e' in the noun "father," which takes the form "отец" in the nominative singular and "отцу" in the dative singular (or the noun "ice," taking the forms "лёд" [nominative singular] and "льду" [dative singular]). Table 1 shows another example of the feminine noun "sister," written as "сестра" in the nominative singular and as "сестёр" in the genitive plural. Variations in stem spelling are however not as important as in other languages, such as Finnish. Finally, to complete our description, we should mention that a limited number of nouns, mainly those borrowed from other languages, are not declined.

Inflectional suffixes may also be attached to particles, numerals, and adjectives. According to Russian grammar rules, adjectives agree in gender, number, and case with the noun they modify. The adjective forms may be one of two major types: long adjectives, inflected for case, gender, and number (e.g., as in "John put the red hat"), and the short form, existing only in the nominative predicate form (e.g., "the hat is red") and inflect only for gender and number. In Russian, indeclinable forms include adverbs, prepositions, conjunctions, plus a limited number of borrowed substantives.

In our experiments we make use of "light" stemmers that apply 57 rules in order to remove only the inflectional suffixes from nouns and adjectives (to normalize the resulting stems, we added four more rules).

Suffixes may also be used to derive new words from a stem, usually by changing the word's part of speech (e.g., "care" and "careful" or "carefulness"). Primarily, Russian derivations are formed through the use of prefixes and suffixes (e.g., "спрутник" [spoutnik] = "с" [prefix] + "пут" [stem, path] + "ник" [suffix]). Forming these words is not always simple, especially without modifying the base form, as in "admit" and "admittance." Just as with English words, Russian consonants and vowels may be shifted, mutated, or dropped. The root serves as the derivation's base and center, and it may or may not occur without the use of word-formative components. In developing aggressive stemmers we concentrated primarily on removing adjectival qualitative and relational suffixes (e.g., "кровь" [blood] and "кровавый" [bloody]). We thus completely ignored any prefixes we thought might change the base word's meaning (e.g., "prehistory" and "history"). Their removal may end up with a base form having unrelated or no meaning, thus diminishing retrieval performance (e.g., "закат" means "sunset" but it could be erroneously interpreted as "за" + "кат" where "за" means "after, behind," and "кат" means "kath" [*Catha edulis*], bushman's tea). To develop our light stemmer and to remove certain derivational suffixes, 40 rules were added to the light stemmer version.

Compound word construction (e.g., handgun, viewfinder) is another morphological characteristic that might impact retrieval effectiveness. Most European languages use some form of compound construction, indicated either by a hyphen (e.g., in French "porte-clefs" [key ring]) or by a suffix attached to the genitive case (e.g., in German with the "-s" suffix in "Produktionsmethode" = "Produktion" + "-s" + "Methode"). In general, however, no particular "glue" is used to build a compound from two or more words, as in English ("viewpoint") or German ("Bankgesellschaft"). Compound constructions are also possible in Finnish, such as "rakkauskirje" = "rakkaus" (love) and "kirje" (letter). In Russian also, frequently encountered word forms include "радиоприёмник" (radio-receiver) = "радио" (radio) + "приёмник" (receiver), or "микроволновой" (adjective) = "микро" (micro) + "волновой" (wave) (with "волновой" = "волна" (stem, noun wave) + "ов" (suffix used to form an adjective from a noun) + "ой" (inflection denoting the masculine, nominative, singular case).

In our efforts to improve pertinent matches between topics and documents written in Russian, we have also created a stopword list, which includes 412 most commonly used terms such as pronouns (e.g., "мы" [we]), prepositions (e.g., "в" [in], "на" [on]), conjunctions (e.g., "и" [and], "или" [or]), or other forms (e.g., "да" [yes], "буду" [will]), etc. Both our stemmers and stopword lists are freely available (http://www.unine.ch/info/clef/).

Test Collection

The Russian test collection used in our experiments was built during the domain-specific tracks at CLEF 2005 to 2008. The main objective of this track is to evaluate the relative performance of various retrieval models for structured scientific bibliographic collections written in English, German, and Russian. In this case, documents contain textual elements (title, abstracts) as well as subject keywords from controlled vocabularies. The main focus is on the evaluation of IR models with a short description of information items, on the one hand, and on the other the leveraging of controlled vocabularies and other structured metadata entities to hopefully improve monolingual and bilingual information retrieval.

From this test suite we extracted the Russian test collection consisting of the Russian Social science corpus (RSSC) comprising 94,581 documents, and the INION corpus covering Russian social science and economics bibliographic data (145,802 articles). Document length in each corpus is rather short, being 19 and 15 distinct indexing terms, respectively. Some statistics about this test collection are given in Table 2.

TABLE 2. Test collections statistics (CLEF).

	2005	2006	2007	2008
Source	RSSC	RSSC INION	INION	INION
Size	64.6 MB	145.5 MB	80.9 MB	80.9 MB
Number of documents	94,581	240,383	145,802	145,802
Number of topics	25	25	25	25
Topics	#126–#150	#151–#175	#176–#200	#201–#225

Typical documents from each collection are listed in Table 3 (RSSC corpus) and Table 4 (INION corpus). To build document representatives during the indexing process, we retained pertinent sections only. These included the ⟨TITLE⟩ and ⟨TEXT⟩ for the RSSC collection and the ⟨TITLE-RU⟩, ⟨KEYWORD-RU⟩ (terms extracted manually from INION Thesaurus) and ⟨ABSTRACT-RU⟩ segments (available for around 27% of the documents) in the INION corpus. In our experiments we ignored additional information such as author name (⟨AUTHOR⟩ or ⟨AUTHOR-RU⟩) or classification tags (e.g., ⟨HANDLE⟩).

Created for domain-specific tasks during CLEF campaigns held during the years 2005–2008, the test collection contains 100 topics. The relevance judgments were made by human assessors, and for six topics no relevant document could be found, leaving 94 topics for the evaluation. These topics covered various subjects (e.g., "Health risks at work," "Doping and sports," "Value change in Eastern Europe"), including both regional ("The German school system") and international topics ("Poverty").

Based on the TREC model, each topic was divided into three logical sections. First we can find a brief title (under the tag ⟨EN-TITLE⟩ in Table 5) followed by a one-sentence description (e.g., ⟨EN-DESC⟩ in Table 5) and a narrative part specifying the relevance assessment criteria (e.g., ⟨EN-NARR⟩ in Table 5). Full examples written in the Russian and English languages are depicted in Table 5. In order to more closely reflect queries sent to commercial search engines in our experiments we used only the title part of the topic formulations. When using only the title section, our queries had a mean size of 3.25 search terms.

IR Models

In order to obtain a broader perspective on the relative merit of the various retrieval models and stemming approaches, we applied two vector-space schemes and three probabilistic models. First we adopted the classical *tf idf* model, wherein the weight attached to each indexing term was the product of its term occurrence frequency (or tf_{ij} for indexing term t_j in document d_i) and its inverse document frequency (or idf_j). To measure similarities between documents and requests, after normalizing (cosine) the indexing weights we computed the inner product (for more information, see Chapter 6 in Manning et al., 2008).

For the vector-space model better weighting schemes have been suggested, especially in cases where the occurrence of a term in a document is viewed as a rare event. Thus, a good practice may be to give more importance to the first occurrence of a term, as compared to its successive and repeating occurrences, with the *tf* component being computed as the $\ln(tf)+1$ or as $\ln(\ln(tf)+1)+1$. A term's presence in a shorter document might also provide stronger evidence than it would in a longer document. In order to take document length into account, we could make use of more complex IR models, including the "*dtu-dtn*" IR model suggested by Singhal, Choi, Hindle, Lewis, and Pereira (1999). In this case Equation 1 calculates the indexing weight assigned to document

TABLE 3. Example of document from RSSC collection.

<DOCNO> RSSC-SOCIONET-RU-20050228-001018 </DOCNO>
<HANDLE> RePEc:rus:cemicf:704 </HANDLE>
<AUTHOR> Зеликина Л.Ф.; Зеликин М.И. </AUTHOR>
<TITLE>
Многофакторные модели экономического роста переходного периода структуры магистральных многообразий. </TITLE>
<CLASSIFICATION>
экономика </CLASSIFICATION>
<TEXT>
Тез. IV Международного семинара ''Комплекс ные исследования'' перехода России и других стран к устойчивому развитию с использованием математического моделирования'' -- Москва, Ин-т социально-политических, исследований РАН сентябрь 1998. </TEXT>

TABLE 4. Example of article extracted from INION corpus.

<DOCNO> ISISS-RAS-ECOSOC-20060324-45953 </DOCNO>
<AUTHOR-RU> Орлов, Г.М.; Кондратенко, А.И. </AUTHOR-RU>
<TITLE-RU>
Социальное партнерство или усиление экономической зависимости редакционных коллективов </TITLE-RU>
<KEYWORDS-RU>
пресса; социальное партнерство; Россия </KEYWORDS-RU>
<ABSTRACT-RU>
По данным анализа деятельности редакционного коллектива газеты ''Орловская правда''. </ABSTRACT-RU>

TABLE 5. Example of topic description in English and Russian languages.

<TOP>
<NUM> 160 </NUM>
<EN-TITLE> Precarious working conditions </EN-TITLE>
<EN-DESC> Research papers and publications on types of work that deviate from normal working conditions </EN-DESC>
<EN-NARR> What "atypical" types of work conditions have developed? What "precarious" consequences are there for effected workers? What improvements to healthcare, social security and unemployment insurance status are being discussed? Are there factors that may halt this development? </EN-NARR> </TOP>
...
<TOP>
<NUM> 160 </NUM>
<RU-TITLE> Опасные условия труда </RU-TITLE>
<RU-DESC> Найти научные статьи и публикации в прессе о видах работ, при которых условия труда отличаются от нормальных. </RU-DESC>
<RU-NARR> Какие существуют «нетипичные» виды условий труда? Каковы могут быть опасные последствия для работников? Какие обсуждаются возможные улучшения в здравоохранении, социальном обеспечении и страховании от безработицы? Существуют ли факторы, способные остановить развитие ситуации? </RU-NARR>
</TOP>

term (*dtu*) and Equation 2 the indexing weight assigned to query term (*dtn*):

$$\begin{aligned} w_{ij} = &[[\ln(\ln(tf_{ij}) + 1) + 1] \cdot idf_j]/ \\ &[(1 - slope) \cdot pivot + (slope \cdot nt_i)] \end{aligned} \quad (1)$$

$$w_{qj} = [[\ln(\ln(tf_{qj}) + 1) + 1] \cdot idf_j] \quad (2)$$

where nt_i is the number of distinct indexing term in document d_i and *pivot* and *slope* are used for adjusting term weight normalization value according to document length. This formulation prevents the retrieval system from overfavoring short documents compared to articles longer than the mean corresponding to the pivot value. For all our experiments the constant slope was fixed at 0.25 and pivot at 15 corresponding to the average document length.

In addition to these two vector-space schemes, we also considered Okapi probabilistic models (Robertson, Walker, & Beaulieu, 2000), as well as two models derived from the Divergence from Randomness (DFR) paradigm (Amati & van Rijsbergen, 2002) wherein the two information measures formulated below were combined:

$$w_{ij} = Inf^1{}_{ij} \cdot Inf^2{}_{ij} = -\log_2[Prob^1{}_{ij}] \cdot (1 - Prob^2{}_{ij}) \quad (3)$$

where $Prob^1{}_{ij}$ is the pure chance probability of finding tf_{ij} occurrences of the term t_j in document d_i. On the other hand, $Prob^2{}_{ij}$ is the probability of encountering a new occurrence of term t_j in the document, given that tf_{ij} occurrences of this term had already been found. To estimate these probabilities, we might instead use the DFR-GL2 model based on the following formulae:

$$Prob^1{}_{ij} = [1/(1 + \lambda_j)] \cdot [\lambda_j/(1 + \lambda_j)]^{tfij} \text{ with } \lambda_j = tc_j/n \quad (4)$$

$$\begin{aligned} Prob^2{}_{ij} &= tfn_{ij}/(tfn_{ij} + 1) \text{ with } tfn_{ij} \\ &= tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl)/l_i)] \end{aligned} \quad (5)$$

where tc_j is the number of occurrences of term t_j in the collection, n the number of documents in the corpus, l_i the length of document d_i, *mean dl* (fixed at 15) the average document length, and c a constant (fixed empirically at 1.5).

In our second DFR model, DFR-I(n_e)B2, Equation 6 is used to calculate $Inf^1{}_{ij}$, and Equation 7 to calculate $Prob^2{}_{ij}$, as shown below:

$$\begin{aligned} & Inf^1{}_{ij} = tfn_{ij} \cdot \log_2[(n + 1)/(ne + 0,5)] \\ \text{with} \quad & ne = n \cdot [1 - [(n - 1)/n]tc_j] \\ \text{and} \quad & tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl)/l_i)] \end{aligned} \quad (6)$$

$$Prob^2_{ij} = 1 - [(tc_j + 1)/(df_j \cdot (tfn_{ij} + 1))] \quad (7)$$

Finally, we also considered an approach based on a language model (LM) (Hiemstra, 2000), known as a non-parametric probabilistic model. Probability estimates would not be based on any known distribution (as in Equation 4), but rather be estimated directly and based on occurrence frequencies in document d_i or the entire C corpus. Within this language model paradigm, various implementations and smoothing methods (Zhai & Lafferty, 2004) might also be considered, and in this study we adopted a model proposed by Hiemstra (2000) as described in Equation 8, which combines an estimate based on document ($P[t_j|d_i]$) and corpus ($P[t_j|C]$):

$$\begin{aligned} & P[d_i|q] = P[d_i] \cdot \Pi_{t_j \in Q}[\lambda_j \cdot P[t_j|d_i] + (1 - \lambda_j) \cdot P[t_j|C]] \\ \text{with} \quad & P[t_j|d_i] = tf_{ij}/l_i \text{ and } P[t_j|C] = df_j/lc \\ \text{with} \quad & lc = \Sigma_k df_k \end{aligned} \quad (8)$$

where λ_j is a smoothing factor (fixed at 0.25 for all indexing terms t_j), df_j indicates the number of documents indexed with the terms t_j, and lc are constants related to the underlying corpus C.

In Equation 8, $P[d_i]$ is the previously calculated probability that the document d_i is pertinent. We ignored this value

in our experiments because it did not vary across the documents and thus did not change the final ranking. For Web searches, however, this probability may vary across different Web pages, depending on the number of incoming links, page lengths, or other factors such as page popularity measures within the Website (Kraaij, Westerveld, & Hiemstra, 2002).

Evaluation

To evaluate the retrieval performance of the various IR schemes we used the mean average precision (MAP), a performance measure that has been used by all evaluation campaigns for more than 15 years in order to objectively compare various IR strategies, particularly regarding their ability to retrieve relevant items (ad hoc tasks) (Buckley & Voorhees, 2005). The MAP value does not have a direct interpretation for the final user. It is computed as the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved (Buckley & Voorhees, 2000). The MAP values were computed by TREC_EVAL software, based on a maximum of 1,000 retrieved records. By using a mean to measure performance we give equal importance to all queries. We also combined the topic descriptions from the 2005 to 2008 CLEF evaluation campaigns in order to base our results on relatively large number of topics (94 in this case), believing that it is important to perform experiments involving the largest possible number of observations.

In order to statistically determine whether one strategy was better than another, we used the two-sided t-test (Buckley & Voorhees, 2005), with the null hypothesis H_0 stating that both retrieval strategies produce a similar MAP. This null hypothesis is accepted if two retrieval schemes returned statistically similar MAP, otherwise it is rejected. In the experiments presented in this paper, statistically significant differences were detected by a two-sided t-test with a significance level of $\alpha = 5\%$.

Finally, it is also well known that the basis for comparisons between two (or more) IR strategies must be similar, using the same document collection and the same topics, as mentioned by (Buckley & Voorhees, 2005):

> "The primary consequence of the noise is the fact that evaluation scores computed from a test collection are *relative* scores only. The only valid use for such scores is to compare them to scores computed for other runs using the exact same collection." (Buckley & Voorhees, 2005, p. 73).

Thus, it is clearly impossible to compare the performance obtained using a test collection with that achieved based on another document collection or directly performances obtained from the CLEF 2007 topics with those of CLEF 2008.

IR Models Evaluation

Table 6 depicts the MAP based on the methodology mentioned above and using four different stemming approaches and six IR models. The last column lists a 4-gram language-independent indexing approach (McNamee & Mayfield, 2004). In this indexing scheme words are decomposed by overlapping 4 letter sequences (the value 4 was selected because it produced the best IR performance). For example, the sequence "prime minister" generates the following 4-grams {"prim," "rime," "mini," "inis," $\cdots$ and "ster"}.

TABLE 6. MAP of various stemming strategies and IR models.

	MAP (Mean Average Precision)				
	None	Light	Aggressive	Snowball	4-gram
tf idf	0.0739*	0.1302*	0.1328*	0.1282*	0.1381*
dtu-dtn	**0.0999**	**0.1892**	0.1749	**0.1847**	0.1708
Okapi	0.0881*	0.1734	0.1735	0.1648	0.1710
DFR-I(n_e)B2	0.0928	0.1802	**0.1812**	0.1734	**0.1741**
DFR-GL2	0.0879*	0.1708	0.1688	0.1624	0.1712
LM	0.0964*	0.1821	0.1793	0.1762	0.1613*
mean	0.0898	0.1710	0.1684	0.1650	0.1644
% change		+90.3%	+87.5%	+83.6%	+83.0%

In Table 6 the best performance obtained for each stemming approach is shown in bold, indicating that either the vector-space model *dtu-dtn* or the probabilistic model DFR-I(n_e)B2 would always prove to be the best IR model. We then used these best performances as a baseline for statistical testing. Any performance differences that were statistically significant when compared to the best IR model are indicated with an asterisk. We can thus see that when compared to *tf idf*, the differences were always statistically significant. For the other models the performance differences were usually not statistically significant (except for the LM model with the 4-gram indexing or those listed in the "None" column).

In addition to the indexing strategies shown in Table 6, we also tested one where stemming was combined with a decompounding procedure. Even though decompounding may be effective for some languages (e.g., German and Finnish), for Russian it resulted in a lower MAP than it would have in the strategy not using decompounding (around 5% in average).

Finally, for all experiments listed in Table 6 we used our stopword list to remove very frequent and noncontent-bearing terms. We also compared retrieval effectiveness of different IR models with and without this list, discovering that performance differences were rather small (around 2% on average), thus showing no evidence that removing the stopword list had any important impact on the MAP.

Stemming Strategies Evaluation

As shown in Table 6, we first evaluated the retrieval performance without any stemmer, listing the MAP values in the "None" column. We then reported the retrieval performance obtained by our "Light" and "Aggressive" stemmers. In the "Snowball," column we listed the MAP obtained using the available Snowball stemmer (http://snowball.tartarus.org/) and in the last column we listed the results of the language-independent 4-gram indexing strategies. As shown

in the second to last row of Table 6, we computed the average performance achieved by each of the six retrieval models in order to obtain an overview of the performance of each stemming approach.

As shown by the values listed in Table 6, all approaches using stemming performed much more effectively than those that did not use stemming. When compared to an approach without stemming (the "None" column in Table 6), averaging the performance over six given models showed relative increases ranging from 83% with the 4-gram indexing scheme to 90.3% with our "light" stemmer (percentage depicted in the last row of Table 6). These relative improvements were clearly quite large and more significant than those found for other European languages (e.g., +4% with the English language, +4.1% Dutch, +7% Spanish, +9% French, +15% Italian, +19% German, +29% Swedish, or +40% Finnish; Tomlinson, 2004).

After applying our statistical tests we found that the performance differences between stemming and no stemming schemes were always statistically significant. Finally, when comparing the 4-gram to the word-based indexing strategies (other than those listed under "None"), performance differences were rather small (e.g., −3.8% over the "light" stemmer), and these performance differences were never statistically significant.

To analyze the effect of applying a stemming, we performed a query-by-query analysis, concentrating only on a single retrieval model DFR-I(n_e)B2, one of the best-performing models for any of the indexing strategies used. In this study we thus showed that applying a stemmer could increase the performance for more than 60 topics (61 with "light," 67 with "aggressive") over a no stemming scheme, and in both cases it was observed that a decrease occurred in average precision (AP) for only 18 topics.

When using the light stemmer the greatest improvement was obtained by Topic #223 "Media in the preschool age," with an AP of 0.6607 compared to 0.01 without stemming. This improvement can be explained by the fact that the term "детьми" (children, instrumental) is found in the topic while terms "детей" (children, accusative or getitive) or "дети" (children, nominative) can be found in the relevant documents. These variants are conflated to the same stem with both our light or aggressive stemmers, but do not with Snowball stemmer (AP of 0.0227), nor do they yield the same 4-gram (AP: 0.0598).

We found a somewhat similar situation with Topic #160 "Precarious working conditions" when the terms "опасные" and "опасньх" were conflated to the same stem and significantly improved the performance for all stemming procedures (e.g., AP of 0.0093 with "None" vs. 0.6165 with "Light"). At times, of course, stemming can diminish retrieval performance, usually through conflating nonrelated terms into the same stem.

We also found that Topic #146 was the most difficult topic in this "Diabetes Mellitus" ("Диабет меллитус") collection. It did not retrieve any items, relevant or not, since none of the terms in the topic appeared in the collection.

Conclusion

In this paper we have presented the main aspects of Russian morphology and suggested two stemmers for this Slavic language, one removing only inflectional suffixes (denoted "light") and a second removing certain frequent derivational suffixes (denoted "aggressive"). Both approaches apply a few rules to correct orthographic irregularities. We have also suggested a stopword list containing 412 word forms. These linguistic tools are freely available on the Internet (www.unine.ch/info/clef/).

To evaluate our stemming approaches we use the most effective current IR models, finding that those IR models derived from the DFR paradigm or the vector-space model *dtu-dtn* perform best, depending of the underlying indexing and stemming strategy. Statistically speaking, these approaches perform better than the classical *tf idf* model or in some cases than a language model, while for the Okapi model there are no significant statistical differences.

When applied to the Russian language, our various experiments clearly show that a stemming procedure improves retrieval effectiveness, especially in the case of the collection containing short documents (e.g., bibliographical records, table or picture captions, statistical tables, etc.). From a statistical point of view, the differences are always significant when compared to an approach ignoring stemming. When comparing different stemming strategies, for most IR models we observe that even though our light stemming tends to perform better than other stemming strategies, performance differences among these different stemmers are never statistically significant. Based on our various examples, we also show that stemming can have a concrete effect on various topic formulations.

In our opinion, when comparing stemming procedures it is also important to consider the final user. A non-stemming or a light stemming approach is better understood than a more aggressive approach that might return unexpected results. For this same reason, for the Russian language we suggest applying a light stemmer, only removing the plural and grammatical cases associated with nouns or adjectives.

Acknowledgment

This research was supported in part by the Swiss NSF under Grant #200021-113273.

References

Amati, G., & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM-Transactions on Information Systems, 20, 357–389.

Buckley, C., & Voorhees, E.M. (2000). Evaluating evaluation measure stability. In Proceedings of 23rd ACM Special Interest Group on Information Retrieval (ACM SIGIR) (pp. 33–40). New York: ACM Press.

Buckley, C., & Voorhees, E.M. (2005). Retrieval system evaluation. In E.M. Voorhees, D.K. Harman (Eds.), TREC Experiment and Evaluation in Information Retrieval (pp. 53–78). Cambridge, MA: MIT Press.

Hiemstra, D. (2000). Using language models for information retrieval. CTIT Ph.D. Thesis.

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of Finnish text documents. In Proceedings of the Ninth ACM Conference on Information and Knowledge Management (ACM-CIKM) (pp. 625–633). New York: ACM Press.

Kraaij, W., Westerveld, T., & Hiemstra, D. (2002). The importance of prior probabilities for entry page search. In Proceedings of the 25th Special Interest Group on Information Retrieval (ACM SIGIR) (pp. 27–34). New York: ACM Press.

Lovins, J.B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11, 22–31.

Malherbe, M. (1995). Les langues de l'humanité. Paris: Robert Laffont.

Manning, C., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, UK: Cambridge University Press.

McNamee, P., & Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. IR Journal, 7, 73–97.

Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., & Stempfhuber, M. (Eds). (2007). Evaluation of multilingual and multi-modal information retrieval. Lecture Notes in Computer Science, 4730. Berlin, Germany: Springer.

Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., & Santos, D. (Eds). (2008). Advances in multilingual and multimodal information retrieval. Lecture Notes in Computer Science, 5152. Berlin, Germany: Springer.

Petras, V., Baerisch, S., & Stempfhuber, M. (2008). The domain-specific track at CLEF 2007. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D.W. Oard, A. Peñas, & D. Santos (Eds.), Advances in multilingual and multimodal information retrieval (pp. 160–173). Lecture Notes in Computer Science, 5152. Berlin, Germany: Springer.

Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14, 130–137.

Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. Information Processing & Management, 36, 95–108.

Savoy, J. (1993). Stemming of French words based on grammatical category. Journal of the American Society for Information Science, 44, 1–9.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. Information Processing & Management, 33, 495–512.

Savoy, J. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages. In Proceedings of the 21st ACM Symposium on Applied Computing (ACM SAC), (pp. 1031–1035). New York: ACM Press.

Savoy J. (2007). Searching strategies for the Bulgarian language. Information Retrieval, 10, 509–529.

Singhal, A., Choi, J., Hindle, D., Lewis, D.D., & Pereira, F. (1999). AT& T at TREC-7. In Proceedings of Seventh Text Retrieval Conference (TREC-7) (pp. 239–251). Gaithersburg, MD: NIST.

Sproat, R. (1992). Morphology and computation. Cambridge, MA: MIT Press.

Tomlinson, S. (2004). Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003. In C. Peters, J. Gonzalo, M. Braschler, M. Kluck (Eds.), Comparative evaluation of multilingual information access systems (pp. 286–300). Lecture Notes in Computer Science, 3237. Berlin, Germany: Springer.

Xu, J., & Croft, B. (1998). Corpus-based stemming using cooccurrence of word variants. ACM-Transactions on Information Systems, 16, 61–81.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems, 22, 179–214.