The Polish Task within Cultural Heritage in CLEF (CHiC) 2013. Torun Runs

Piotr Malak[1, 2]

[1] Institute of Information Science and Book Studies, Nicolaus Copernicus University, Torun, Poland
piomk@uni.torun.pl
[2] Department of Computer Science, University of Neuchatel, Neuchatel, Switzerland
Piotr.Malak@unine.ch

Abstract. This paper presents the goals and realization of a Polish Task for Cultural Heritage in CLEF (CHiC) 2013 campaign. We present a short introduction to Polish language complexity, and problematic issues that may occur during automatic text processing. The organization of a separate ad-hoc task for Polish has been described, as well as collection used for test. We do also describe topics delivered for the task. Last part of the paper presents results analysis and comments on used IR techniques and their adequacy for information retrieval for Polish.

Keywords. CHiC Polish Task, ad-hoc IR, OKAPI, tf/idf

1 Introduction

Cultural heritage preservation in digital form is additional, and sometimes more secure than physical one. It give the possibility of at least seeing the image of CH object via Internet, if not possible to see original. EUROPEANA is promising example of digital library focusing on CH objects [9]. Its goal is to deliver millions of European cultural heritage objects to a wide public. The CH objects available there are delivered by different institution all over Europe. Thus different data and meta data formats are being used for documents. Also there is a variety of media types being represented in Europeana resources, as well as different languages in documents descriptions. Making such w rich set of objects, objects descriptions and formats types easily available to the users requires a sophisticated approach. ChiC (Cultural Heritage in CLEF) evaluation lab attempts to provide evaluation of cultural heritage digital resources [1].

Among subtasks of CHiC there is one new – *Polish Task*, devoted to Polish objects in Europeana resources. The main objective of Polish task is to increase knowledge of different IR issues for handling complex language, such as Polish, with a special attention to CH objects. Two kinds of actions have been allowed for this subtask: auto-

matic one, and manual enrichment [9, 7]. Torun has participated in both, however only the manual enrichments have been submitted.

During this evaluation lab we want to answer following hypothesis:

1. Can we assume for flexional languages, and particularly for Polish, that morphological complexity has none, or relatively small, impact on the retrieval performance?
2. Will use of a light stemmer improve the searching efficiency?

For answering those questions we deliver a collection of Polish documents from Europeana resources, as well as set of 50 queries to be examined on the collection.

The rest of this paper is organized as follow: section 2 presents short introduction to Polish morphological complexity, as well as other language related issues, typical for CH objects handling. Section 3 describes experiment setup, its requirements and organization. This is followed by section 4, devoted to results and their analyses. Finally section 5 concludes the experiment.

2 Morphological complexity

Polish itself is a challenging language for IR, because of its morphological complexity. For example, one can distinguish 11 main classes of verb conjugation, which are not always regular. Also declension offers quite a lot of irregularities. Relatively free word order causes some difficulties in automatic POS tagging. As for English documents on Polish language, and its grammar one may refer to [4, 14, 6], here we will give only short presentation of possible problems for IR in Polish.

2.1 Nouns

A noun declension is being ruled by seven cases, which offer seventeen declension types. There are following cases in Polish: nominative (*mianownik*), genitive (*dopełniacz*), dative (*celownik*), accusative (*biernik*), instrumental (*narzędnik*), locative (*miejscownik*), vocative (*wołacz*).
For nouns there are two number classes, and three main genders: masculine (with subclasses: personal in sing., non-personal animate, non-personal inanimate) feminine, and neutral. All the classes distinguish by proper suffix, but one may not forget, that in many cases not only suffix, but also a words stem (root) derives. Quite often one may meet noun with different stem for different case, like for *człowiek* (*a man*) or *kolega* (*colleague*). Both nouns are masculine, personal.

Table 1. Irregular declension of noun *człowiek* (a man)

	singular	plural
N.	człowiek	ludzi**e**
G.	człowieka	ludzi
D.	człowiek**owi**	ludzi**om**
A.	człowieka	ludzi
I.	człowiek**iem**	ludź**mi**
L.	człowiek**u**	ludzi**ach**
V.	człowiek**u!**	ludzi**e!**

Table 2. Irregular declension of noun *kolega* (a colleague)

	singular	plural
N.	koleg**a**	koledz**y**
G.	koleg**i**	koleg**ów**
D.	koledz**e**	koleg**om**
A.	koleg**ę**	koleg**ów**
I.	koleg**ą**	koleg**ami**
L.	koledz**e**	koleg**ach**
V.	koleg**o**!	koledz**y**!

Also, unlike in most of the languages, personal names are subject of declension in Polish, and, furthermore, in Polish language foreign personal names, are also declined. As example form our topics we may call *Maria Skłodowska-Curie* (Maire Curie) or *Fryderyk Szopen* (Frédéric Chopin). During declension both name, and last name are being subject of suffix changes. This linguistic feature is also present for other names, such as geographical (e.g., Gibraltar, -u, -owi, -, -rem, -rze, -rze)

2.2 Verbs

Most of verbs are of regular conjugation, but there are also irregularities in verb conjugation, like for *iść* (to walk).

Table 3. Sample conjugation of irregular verb *iść* (to walk)

person	**present simple**	**past simple m., / f. / n.**
1 s.	id**ę**	szedłem / szłam / szłom
2 s.	idziesz	szedłeś / szłaś / szłoś
3 s.	idz**ie**	szedł / szła / szło
1 pl.	idziemy	szliśmy / szłyśmy / szłyśmy
2 pl.	idziecie	szliście / szłyście / szłyście
3 pl.	id**ą**	szli / szły / szły

2.3 Spelling changes

Another problem for specific task of IR for cultural heritage objects are changes that took place in the language over the time. Some changes concern meaning of the words, and some concern notations. As an example of meaning change we may give a word *rzeźba* currently meaning a sculpture. But its historical meaning was also a slaughter, a massacre [12]. As for example of changes in spelling one may consider word *sejm* (parliament), which used to be written as *seym* in 18th and 19th centuries, as well as in the beginning of 20th century. The original notation occur in CH documents titles, but, the metatags contain current version, thus there was no need to put additional dictionary of different notations of words. However, for more general purposes on information retrieval also from digitalized historical documents, one should consider preparing appropriate dictionary of notations, as well as meaning (historical synonymy).

2.4 Alphabet

Another problem one may encounter processing Polish texts is notation normalization. There are characteristic letter as: ą, ć, ę, ł, ń, ó, ś, ź, ż, ch, dź, dż, rz. There is unified transcription rules set for Polish. However, the Europeana materials, we have been dealing with, are in Polish with original notation kept, so also for the topics we have applied original Polish notation without any normalization.

3 Experiment setup

3.1 Collection

The Polish collection is a part of CHiC 2012, and 2013 Multilingual collection. In total there are 1,093,705 documents in 1,094 files in Europeana's Polish collection. It is the 9th the most reach collection of all 30 languages. The whole collection archive, of a total size of 119 MB. The archive have been made available by Europeana last year, at http://ims.dei.unipd.it/data/chic/. According to CHiC 2012 evaluation [8], Polish collection consists of: 975,818 text documents, 117,075 images, 582 videos, 230 sound documents.

Analysis of collection structure presents Table 4.

Table 4. Structure of CHiC 2013 Polish collection

Media type	documents	percentage of collection
text	975,818	89.221%
images	117,075	10.704%
videos	582	0.053%
sound	230	0.021%

All the collection files are provided in XML schema, as on example.

Fig. 1. Example record from Europeana Polish collection

```
<ims:metadata
ims:identifier="http://www.europeana.eu/resolve/record/09
316/67EC33BB285087B51E5E57DCA56CC39755E5D12C"
ims:namespace="http://www.europeana.eu/"
ims:language="pol">
<ims:fields>
<dc:contributor>Jewish Historical Institute, Warsaw [Col-
lection]</dc:contributor>
<dc:creator>Juedische Gemeinde zu Breslau [Crea-
te]</dc:creator>
<dc:description>Korespondencja, rachunki, plany budowy
dot. przebudowy synagogi przy przytuł ku na mieszka-
nia.</dc:description>
<dc:description>Old call number:</dc:description>
<dc:description>Aufbewahrung/Standort: The Emanuel Rin-
gelblum Jewish Historical Institute War-
saw</dc:description>
<dc:identifier>local JHI-105_1160 [Metada-
ta]</dc:identifier>
<dc:rights>Jewish Historical Institute, War-
saw</dc:rights>
<dc:source>Jewish Historical Institute, War-
saw</dc:source>
<dc:subject>18th - 20th century</dc:subject>
<dc:subject>History of Jewish Community in Wroclaw, Po-
land</dc:subject>
<dc:subject>Handelsgenehmigung fur Juden</dc:subject>
<dc:subject>Wroc ł aw, Poland</dc:subject>
<dc:title>Korrespondenz, Rechnungen, Baupläne betreffend
Umbau der Synagoge des Zufluchsthauses in Wohnun-
gen.</dc:title>
<dc:type>deutsch [language]</dc:type>
<dcterms:created>1939-1940 [Create]</dcterms:created>
<europeana:country>poland</europeana:country>
<europeana:dataProvider>Jewish Historical Institute, War-
saw</europeana:dataProvider>
<euro-
peana:isShownAt>http://judaicaeuropeana.pl/zbiory.php?id=
105_1160</europeana:isShownAt>
<europeana:language>pl</europeana:language>
```

```
<euro-
peana:object>http://judaicaeuropeana.pl/105_1160/105_1160
_Strona_290.jpg</europeana:object>
<europeana:provider>Judaica Euro-
peana</europeana:provider>
<europeana:rights>http://www.europeana.eu/rights/rr-
f/</europeana:rights>
<europeana:type>TEXT</europeana:type>
<euro-
peana:uri>http://www.europeana.eu/resolve/record/09316/67
EC33BB285087B51E5E57DCA56CC39755E5D12C</europeana:uri>
</ims:fields>
</ims:metadata>
```

As described in [8], Europaena data *are metadata describing digital representations of cultural heritage objects*. As within those data one may find different schemas, like:

- Dublin Core (all tags starting with *dc:* prefix),
- Qualified Dublin Core (all tags starting with *dcterms:* prefix), and
- Europeana Semantic Elements (tags with *europeana:* prefix).

To make indexing process faster the following set of fields have been included:

- <dc:contributor>, <dc:creator>, <dc:date>, <dc:language>, <dc:subject>, <dc:title>, <dc:type>, <dcterms:alternative>, <dcterms:created>, <europeana:language>, <europeana:type>, <europeana:uri>, <europeana:year>.

3.2 Topics

For Polish task at CHiC 2013 a set of 50 topics have been prepared, to be searched over a test collection. The topics are short or average expressions, all together they consists of 141 word tokens, which gives 2,82 word per topic. There is 10 monogram topics, 11 bigrams, and 29 longer. The longest topics consist of 6 words, we have four such long topics. They consist of two, up to three connectives.

Topics are given in Polish, and additionally they are released in English translation (CHIC-2013-PL-Polish-Topics.xml, CHIC-2013-PL-English-Topics.xml, respectively). An example of topic is shown in Figure 2 for Polish topics, and in Figure 3 for English translation.

Fig. 2. Example of Polish topic for CHiC 2013 Polish Task

```
<?xml version="1.0" encoding="UTF-8"?>
<topics>
<topic lang="pl">
<identifier>CHIC-2013-PL-001</identifier>
<title>meblarstwo polskie</title>
<description>prace poświęcone polskim meblom, polskiemu
meblarstwu</description>
</topic>
```

Fig. 3. Example of English translation for topic for CHiC 2013 Polish Task

```
<?xml version="1.0" encoding="UTF-8"?>
<topics>
<topic lang="en">
<identifier>CHIC-2013-PL-001</identifier>
<title>furniture joinery</title>
<description>works on Polish furniture or Polish furni-
ture joinery</description>
</topic>
```

Each topic is identified by <identifier> tag, while query itself is being provided within <title> tag. For each query additional <description> field has been provided, basing on previous CHiC's experiences. The aim of this additional field is to give the relevance assessors an idea of what subjects were intended to retrieve with a particular topic. As stated in [5], the <description> field must not be used for retrieval purposes.

Our topics have been prepared on the basis of Europeana search logs, as well as deductions on cultural heritage users interests. As this year Poland has celebrated 150th anniversary of January Uprising, few topics related to Polish territories and history within 18th and 19th centuries have been added. There are also topics on certain historical periods, as well as few on temporary issues concerning Poland. The chronological time frames are always additional narrowing of the general topic, like:

- <title>chłopi w 18 lub 19 wieku</title>
- <title>peasants in 18 or 19 century</title>

There are some named entities, concerning mostly persons, but also a geographical or historical ones. Generally topics consists of:

1. Chronological topics:
   (a) 8 topics with time frames given (XVIII or XIX century),

(b) 8 topics concerning particular period of time, like (Barok, *ang. Baroque*, or Dwudziestolecie Międzywojenne, *ang. interval period (1919 - 1939)*)

2. Named entities:
   (a) 12 topics with personal names (generał Józef Bem, *ang. general Jozef Bem*, or Matka Boska, *ang. Our Lady*)
   (b) 6 topics with geographical names (Kraków, *ang. Cracow*, or pałace Lubelszczyzny, *ang. mansions of Lublin Voivodeship*)
   (c) 5 topics with historical names (Powstanie Styczniowe, *ang. January Uprising*, or Barok, ang. *Baroque*).
3. General entities:
   (a) 5 topics on religion or beliefs (diabeł, *ang. Devil*)
   (b) 7 topics on social groups or functions (robotnicy, *ang. workers*)

3.3 Manually enriched topics

As mentioned earlier, from Torun manual enrichment runs have been submitted. For those two runs only topics have been enriched, although it was possible to enrich the CH objects and/or the queries [9].

Two levels of user's experience have been emulated with the respect to general knowledge level for each of type. As Europeana provides specific contents, a cultural heritage, our enrichments aimed two groups of users: educated (in terms of at least colleague), and specialists (in the terms of knowledge of additional information sources, historical contexts, etc.). For educated users simulation we have mostly used synonyms of terms from topics, as well, as some detailed topics. Specialists enrichment was supported by use of encyclopaedias. There have been also detailed topics added, but with full respect to original topic. Statistically, original topic titles consists of 141 tokens (average 2.82 token per topic), while educated enrichment resulted in 303 tokens (av. 6.1 per topic), and specialists given 489 tokens in total (av. 9.78/topic). For those files either <title>, and <enriched> fields have been used during indexing process.

3.4 Indexing strategies

For each enriched file, as well as for the collection documents stop word removing procedure has been applied. The stoplist consists of 304 entries, as for stop terms all their grammatical forms have been included. The lists includes among the others determinants, prepositions, conjunctions, pronouns. The stopwords removal procedure have been applied for manually enriched topics files, as well as for collection itself.

Further on light stemming procedure have been applied for each of two kinds of manually enriched topic files. For the experiment a light stemming [3, 11] has been used affecting only nouns. Some experiments have already proven light stemming sufficient enough for IR purposes in comparison to morphological ones [11].

As weighting scheme for official runs OKAPI (BM25) probabilistic algorithm has been used [10]. For unofficial, automatic run a statistical tf.idf weighting has been applied, together with Boolean topic keywords matching.

3.5 Evaluation

Results have been evaluated according to the following evaluation schemes:

- MAP, P@5, P@10, p-value, GMAP, MFRS.

Finally, as for previous CHiC experiments, MAP (Mean Average Precision) have been applied. For each topic MAP value has been computed for the first 1000 retrieved documents in ranked list.

4 Results and analyses

4.1 Official runs

From Torun, there have been two manual enriched runs submitted, each one with light stemmer, and without any stemming. There has been also one automatic run, but due to late results formatting that one could not be submitted as official one, however we will refer to this run as well. The results on IR using titles and enrichment fields have been slightly worse, then obtained just for titles, using OKAPI. The baseline file, fully automatic, with stop words removal, without stemming, achieved MAP of 0.314. Comparison of the MAP for submitted runs presents Table 5. One can observe the general educated users emulation resulted in better MAP than expert one either for light stemmer (LS suffix in run name) and for no stemming (NO).

Table 5. MAP for official submitted runs.

Run id	**Parameters**	**MAP**	**% of change**
BASE	only <title> field, no stemming	0.3140	n/a
PLTO1EDULS	educated users, light stemmer	**0.2774**	**-11.66%**
PLTO1EDUNO	educated user, no stemmer	0.2724	-13.25%
PLTO2EXPRTLS	expert user, light stemmer	**0.2690**	**-14.33%**
PLTO2EXPRTNO	expert user, no stemmer	0.2709	-13.73%

4.2 Enriched topics analysis

Results for enriched topic files in values of MAP are worse, than those retrieved just for titles. The differences are around -13%, which statistically is significant. Additional remark is that expert enriched topics gave the worse result than those prepared in respect to educated users, either if there were stemming applied or not.

One of the reasons may be too extensive keywords coverage over the test collection in expert emulation file. Those files offered nearly thirteen tokens per topic in

average (2.8 in title, and 9.78 in enrichment). The average keywords number was then higher by 4 than in educated user files (2.8 + 6.1), and by 10 in comparison to titles only. Such overload of distinguish terms lead to retrieve too many documents in respond. As during indexing each of keywords have been treated as a separate one, there obviously have been more total matched documents for queries richer in the terms of keywords number. Bigger retrieved set means less relevant items in it, as there was stable relevant documents number.

Another possibility of difference to base, and educated run could be specific vocabulary used in expert enriched topics. Experts keywords tended to narrow query, what could lead to retrieving documents relevant to the narrower term, but not to the topic title. As for topic #28: *podróże i relacje* (journeys and stories), expert enrichment consists the following Polish trawelers: *Fryderyk Skarbek, Juliusz Słowacki, Stanisła Potocki, Paweł Strzelecki, Ernest Malinowski, Bronisław Malinowski, Ignacy Domeyko, Benedykt Dybowski*. This is quite specific list giving proper names, including famous Polish writer and poet Słowacki, while for educated user we have more general terms: *ekspedycje* (expeditions), *wyprawy* (journeys), *dziennik podróży* (journey diary).

One can also observe, that the best MAP values reached topic file of educated users, with light stemming applied, however, stemming for experts files received the worst MAP value of all. Generally, as mentioned in preceding paragraphs, educated user emulation resulted in better precision of retrieved documents. One of reasons for this state was smaller number of keywords, while keeping more general terms in enrichment. Generality of enriched terms has also influence on the stems produced from keywords. Those results confirms statement, too much keywords decreases efficiency of information retrieval.

4.3 Comparison of enriched and basic topics

Generally the baseline run performed better in the terms of MAP, than enriched. Here we describe the overall MAP for each submitted runs, in comparison to the baseline one.

Better performance of enriched topics.
Not all submitted enrichments provided worse results, than the base one, however. For some topics they got better Average Precision (AP), than for the title-based run. For the best of enriched runs, educated user, with light stemming, there are 22 topics of AP higher than respectively in the baseline run (while 19 for no stemming). The extreme positive difference +34630% (MAP = 0.3463 to MAP = 0.001) in the terms of average precision got topic #29: *Warszawa w 19 wieku w sztuce* (Warsaw in 19 century in art) in run educated, no stemming (respectively +22140% with stemming). For this topics, the following enrichment have been added by educated user: *architektura* (architecture), *dzielnica* (district), Warszawa. The next enriched topic with MAP higher, than a baseline, was #19 *powstania w Królestwie Polskim* (uprising in Kingdom of Poland). For this topic educated user enrichment was: *Królestwo Polskie, Królestwo Kongresowe, powstania, powstanie styczniowe, powstanie listopadowe*

(Kingdom of Poland, Congress Kingdom (of Poland), uprisings, January Uprising, November Uprising). It received +2781% (MAP = 0.6590 to MAP = 0.237). The same topic get also the best MAP as for no stemming used. In the second case, the positive difference was even greater, +2865% (0.6790 to 0.0237).

For expert enrichment, with light stemming there were also 22 topics of performance better than the baseline, in the terms of average precision (while 20 for experts without stemming). Here we encountered even higher positive difference. For topic #32, *kobiety w powstaniach i w wojsku* (uprising or military and women) there is +70625% (MAP = 0.2825 to 0.0004) better AP, than for the baseline run. The same topic performed the best also for expert enrichment, without stemming. In that case it gained +60150% (MAP = 0.2406) better precision than baseline. Enriched terms was: *Emilia Plater, sanitariuszka, łączniczka, agentka, Baska, Iza, Hanka, Organizacja Piątek* (Emilia Plater – famous noblewoman and revolutionary unit leader; nurse; agent; Baska, Iza , Hanka: - nicknames of nurses or soldiers of Warsaw Uprising; "Five" – organisations of five women supporting uprising, and creating new "fives"). Here, there are additional terms closely related to military service or to persons/ groups taking part in Polish uprisings. The next better result for stemmed expert topics, +9579% (MAP 0.4502 to 0.0047), has been reached for topic #34: *obrazy miasta* (city in paintings). And again, for not stemmed enriched topics it was also second the best performing topic, +8917% in comparison to baseline run.

Better performance of baseline topics.

The baseline run provides, general, better performance than any of enriched runs. For educated users enrichment, the biggest difference in the favor of baseline is for topic #30: *Polska i Europa w 18 wieku* (Poland and Europe in 18 century). Only title based indexing performed for this topic +1536% better than enrichment with light stemming (MAP 0.1014 to 0.0066). While as for no stemmed enrichment topic #36: *kult* (cult/worship) was of the biggest difference. A baseline run was of +2149% better (MAP = 0.7136 to 0.0332). For this topic added terms were too general historically (covering periods extending given time frame) or retrieved documents have been related to only one of topic limitations (like only Poland internal affairs).

For stemmed experts enrichment the worst was topic #35: *święci* (saints). For this one the difference was +7140% in the favor of baseline run (MAP = 0.1428 to 0.0020). And in case of no stemming for experts enrichment, the worst results got topic #47: *AGD* (housewares). The difference was 1641% (MAP = 0.9667 to 0.0207).

Comparison of the greatest differences for individual topics performance presents Table 6.

Table 6. The biggest difference in MAP for the same topic

	Baseline	**PlToEduLS**	**PlToEduNO**	**PlToExprtLS**	**PlToExprtNO**	**highest % difference**
#19	0.0237	0.6590	**0.6790**	0.2262	0.1880	2865
#29	0.0010	0.2214	**0.3463**	0.0042	0.0009	34630

#30	**0.1014**	0.0066	0.0347	0.0083	0.0315	1536
#32	0.0004	0.0058	0.0017	**0.2825**	0.2406	70625
#34	0.0047	0.0409	0.0399	**0.4502**	0.4191	9579
#35	**0.1428**	0.0541	0.0106	0.0020	0.0087	7140
#36	**0.7136**	0.0516	0.0332	0.1410	0.3141	2149
#47	**0.9667**	0.2265	0.4020	0.1643	0.0207	1641

4.4 Unofficial automatic run

There were also one automatic run made in Torun, but since it was not submitted to Polish Task, it is considered as unofficial one. For this run the following settings have been used. First of all for a stopwords removal procedure have been applied both for topic file, and for documents collection. A light stemming has been used, and no enrichment. Indexing relayed on Boolean conjunction of topic keywords, which means, only documents where all the topical keywords have been matched, have been retrieved. Weighting schema for this run was td.idf. With such settings the smallest number, only 9 documents have been retrieved for topic #27: *diabeł w sztuce* (Devil in art). As in DIRECT relevance assessment system there are 187 out of 562 documents marked as relevant or partially relevant. This situation may be a result of very strict documents matching, as well as not using any enrichment for the very run.

In the terms of mean average precision, Torun automatic unofficial (TOAutom) run performed better even than baseline official run. The MAP for TOAutom was 0.3484 while 0.3140 for baseline one, which makes it 11% better. Only 12 topics have average precision worse than in the baseline.

For automatic run the best individual precision has been reached for topic #24: *Fryderyk Szopen* (Fryderyk Chopin). MAP for this topic have been calculated on 0.9959 (while 0.1131 for baseline run), which means nearly all documents retrieved in automatic run, have been relevant for this topic. In general 11 topics have MAP value higher than 0.5 (while 14 for baseline run).

There are two topics very weak, in the terms of precision. They are #49, and #41, of MAP 0.002 (0.1885), and 0.004 (0.6162), respectively for TOAutom run, and a baseline one.

5 Conclusions

Considering experiments, and further relevance assessment evaluation, one may conclude unigram indexing strategy, matching documents only to single keyword from topic is not the best choice for structured CH objects. For example, there is a personal name *Jarosław* in Poland, as well, as a city name. Thus for topic #031: *Lech lub Jarosław Kaczyński*, there were a lot of false positive retrievals, for most of retrieved documents considered the city. This topic was of the worst relevance ratio – for 731 assessed documents only 16 items (2%) have been considered as relevant or partially relevant. Neither vector-space, nor probabilistic models can impose relevant

retrieval of all keywords from the query. Using a semi-Boolean approach (at least as logical conjunction of query terms) seems to be a good strategy, at least in CH domain. However, further improvements are required, as unofficial automatic run did not performed as well, as expected.

The experiments showed that applying light stemmer for the topics files, and collection increases the performance of retrieval, despite of indexing strategy (statistical tf/idf or probabilistic OKAPI), which was also achieved in other experiments [7]. However, further experiments with the use of an aggressive stemmer should be conducted in order to verify influence of such stemming procedures on relevance of retrieved items.

Polish task experiments [7] have shown using *n*-gram od trunk-*n* does not improve the retrieval performance. So, there appear a question if using lemmatizing, instead of stemming, would increase the relevance of retrieved items. Stemming cuts words to a common stem (root), which is not necessarily a proper grammatical form, additionally this procedure may join different words into one stem. Lemmatizing, despite of higher operational costs, deliver proper grammatical form for majority of document vocabulary. This approach could insure better distinction between words having partially the same spelling. Comparing lemmatizing to stemming for Polish CH object is also a subject for further research.

6 Acknowledgments

This research was supported in part by the Sciex-NMS under Grant POL 11.219 *Information Retrieval and Text Categorization for Polish.*

7 References

1. *CHiC: Cultural Heritage in CLEF*, http://www.promise-noe.eu/chic-2013/home
2. *Europeana Europeana: think culture*, http://europeana.eu/
3. Fautsch C., Savoy J.: *Algorithmic Stemmers or Morphological Analysis: An Evaluation*. JASIST. 60, 1616-1624 (2009)
4. Feldstein Ron F.: *A Concise Polish Grammar,* SEELRC 2001 http://www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=4
5. *Guidelines for participation and submission*, on-line: http://www.promise-noe.eu/chic-2013/guidelines-for-participation-and-submission/polish-task
6. Jagodzinski G.: *A Grammar of the Polish Language*, http://grzegorj.w.interia.pl/gram/en/gram00.html.
7. Petras V., et all: *Cultural Heritage in CLEF (CHiC) 2013*
8. Petras V., et all: *Cultural Heritage in CLEF (CHiC) Overview 2012*, on-line: http://www.clef-initiative.eu/documents/71612/0cadb163-3e32-4f16-a659-b457480c2a29
9. *Polish Track at CLEF 2013*, http://members.unine.ch/jacques.savoy/Polish/
10. Robertson S.: How Okapi Came to TREC, in: Harman D., Voorhees Ellen V.: TREC. Experiment and Evaluation in Information Retrieval, (287-299), The MIT Press (2005)

11. Savoy, J.: *Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages*. Proceedings ACM-SAC, 1031-1035. The ACM Press, (2006)
12. *Słownik języka polskiego XVII i 1. połowy XVIII wieku*, on-line: available on World Wide Net: http://sxvii.pl/index.php?strona=haslo&id_hasla=9516&forma=RZE%C5%B9BA#9516
13. *Słownik poprawnej polszczyzny*. Warszawa: PWN, 1995.
14. Swan Oscar E.: *Polish Grammar in a Nutshell*, http://polish.slavic.pitt.edu/firstyear/nutshell.pdf