ECAI 2010
H. Coelho et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-606-5-529

Automatic free-text-tagging of online news archives

Richárd Farkas[1] and **Gábor Berend** and **István Hegedűs**[2]
András Kárpáti and **Balázs Krich**[3]

Abstract. In this paper, we shall introduce the problem of free-text-tagging of online news archives. From an application point of view, it has many benefits for online news portals and on the other hand, the task has unique characteristics compared to existing approaches for free-text-tagging. We shall describe our system, which was developed for the archive (consisting of 370 thousand articles) of the most visited Hungarian news portal `www.origo.hu`, along with research questions encountered and solved during our task. As the evaluation of tagging is not straightforward at the end of the project the news company manually investigated the tagging of the automatic system which yielded an F-measure of 71.9.

1 Introduction

Free-text-tagging is the task of assigning a few natural language phrases to documents which summarize and semantically represent their content [11]. Tags are useful for organizing, retrieving and linking different contents. Here, we shall introduce our automatic free-text-tagging solution especially designed for online news archives along with the experiences gained on the tagging of the Hungarian [origo] news portal.

[origo] is the most visited news site in Hungary, reaching about 45% of all Internet users in the country. The site `www.origo.hu` was launched in December 1998 and more than 370,000 articles have been published . As a typical general interest portal, [origo] covers a very wide spectrum of topics and themes. Origo assigns tags manually to its published contents since February 2009. Taggers are restricted to textual content, while images attached to the articles inherit tags from the article itself.

The manual tagging system of [origo] is about halfway between free-for-all users' folksonomies [13] and expert information systems that are tagged by information specialists. During system design, Origo considered using free-for-all community user tagging in order to exploit the power of folksonomies such as Wikipedia, del.icio.us and other social tagging systems. However, previous experiences of Origo on free-for-all user tagging were similar to Kipp [6], i.e. users generally want to store more than just the subject of the documents; they want to see their relationship to the objects in different ways, express an emotional connection and assign personal data management information to documents. They often use non-subject tags as well, "*tags which are deliberately excluded from traditional classification systems due to their potentially temporary or task specific nature.*"

Hence, Origo decided to employ the so-called *community-based self-tagging*, where tags are assigned to contents by their creator (journalists and editors), but there is no predefined taxonomy for tags. We should remark here that although the number of tagging users is relatively small (around 50 users), the Origo system is not an expert tagging system. This is due to the fact that users have the freedom to apply tags and its members form a very special user community with significantly different skills, interests and cultural backgrounds. This half-way-position between folksonomies and expert systems seems to be a good compromise.

The principles on how users should tag items are summarized in a guideline. It defines four tag types: `topic`, `person`, `organization` and `location`. At least one tag of type `topic` is mandatory for all articles. Type `person` incorporates names of fictive creatures and animals as well. Having the `location` tags, the place of a news can be visualized on a map and location-based queries can be processed. For entities in general, only the ones should be added as tags that play an important role in the article and whose frequency is more than occasional in the whole archive.

The guideline also states what should be applied as a tag (via typical examples), and what should not be used (slang, metaphors, paraphrases, verbs, adjectives, pronouns, etc.). The document also defines the ideal number of tags for different journalistic genres, makes it clear what kind of expressions are too general as a tag and what kind of expressions are too specific to be applied as tags. The guideline also offers advice on how to differentiate meanings that share the same linguistic expression (e.g.: '*László Kovács the politician*' and '*László Kovács the boxer*'), how to avoid creating several tags for the same meaning (e.g.: *H1N1* or *Swine Flu* or *Influenza A virus*).

2 RELATED WORK

Automatic free-text-tagging has gained attention in recent years. Previous works can be categorized into two approaches, namely tag recommendation (or assignment) and keyphrase extraction. Tag recommendation systems [14, 16] rely mainly on formerly tagged corpora. The key idea behind these approaches is to find similar documents and to assign tags of the manually labeled documents to the unlabeled ones. Autotag [14], the pioneering work of tag recommendation, simply applies standard information retrieval metrics to find similar documents and chooses tags from the nearest ones based on frequency information. Many participants of the ECML PKDD tag recommendation challenges [2, 19] also built their systems on document-similarity-based approaches like this.

Such methods, however, have the disadvantage of exploiting tags assigned by humans that are often inappropriate or inconsistent with the whole document set (i. e. the tag cloud). Moreover, these approaches cannot be adapted to the dynamics of topics, as they are not able to involve new tags (since they operate on a predefined set of tags that have been previously assigned to at least one document).

[1] Hungarian Academy of Sciences, Hungary, email: rfarkas@inf.u-szeged.hu

[2] University of Szeged, Hungary, email: {berendg,hegedusi}@inf.u-szeged.hu

[3] Origo Ldt., Hungary, email: {karpati.andras,krich.balazs}@origo.hu

Another drawback of these methods is that they are heavily domain-dependent, which means that each and every time we would like to use them on a document set, labeled documents are necessary.

The keyphrase extraction approach [10, 21, 23] extracts phrases from just one document that are the most characteristic of the given content. In these approaches keyphrase extraction is regarded as a classification task, in which certain n-grams of a specific document function as keyphrase candidates, and the task is to classify them with respect to whether they are proper keyphrases. This raises the problem of assigning only those kinds of tags that are present in a document, although sometimes tags that are not present can be more informative. Moreover, these tags might not be able to easily distinguish a document from all the others and not form a coherent tag cloud. To overcome the shortcomings of inconsistency, Turney [21] tried to involve Web queries to augment the consistency of tags extracted from documents. Gutwin et al. [4] exploited domain specific knowledge to improve the quality of automatic tagging and most recent methods analyzing the term co-occurence graph [10].

Both tag recommendation and keyphrase extraction requires labeled training documents. However, it would be costly to acquire previously tagged domain-specific corpora for training from such diverse topics that the news archive of [origo] covers. In the news archive tagging task, just a relatively small training dataset (manually tagged news) was available and the coherence of the global tag set through the whole archive was a key objective. Hence, our global solution lies between keyphrase extraction and tag recommendation, as its chief goal is to represent the content of each news item (local information), while the tag set should be consistent; e.g. topics at different levels may use the same tags and one particular phrase should be chosen from a certain set of synonyms (global consistency).

One of the key subproblems which has to be addressed in free-text-tagging is that of recognizing semantic relatedness among terms. The classic approaches for this are co-occurrence-based measures (like Latent Semantic Indexing [9]) and metrics derived from the path between the concepts of a taxonomy (usually from the hypernym tree of Wordnets) [15]. Relatedness calculated from co-occurrences may be noisy, while the coverage of taxonomies are generally low. To overcome these disadvantages, most recent studies have suggested that we should exploit the semi-structured Wikipedia as the source of semantic relatedness information. WikiRelate! [17] – the first published study on this field – used redirections, disambiguation pages and categories of Wikipedia. We shall also introduce five Wikipedia-based metrics for abstraction (which is a special type of semantic relatedness). The work of Grineva et al. [5] is the closest one to our approach. In it they describe how they constructed a graph whose nodes were terms of a document and the weights of edges were derived from the link structure of Wikipedia; and finally, keyphrases of a document were determined by analyzing this graph.

3 AUTOMATIC TAGGING

Our main approach of finding the best set of tags for an article from the archive followed a three-step strategy: (1) the extraction of potential tags from the document itself, based on a linguistic analysis; (2) the extension of the set of potential tags exploiting semantic knowledge obtained from external sources (like Wikipedia) and from the whole corpus; (3) the filtration of the set of potential tags to an appropriate size, based on global statistics.

3.1 Potential tags from the document

In the first step, key concepts being present in the text were gathered as a set of potential tags. These key concepts consist of person-names, organization names and trademarks which are the chief actors of the news, places which can be directly assigned to it and noun phrases which can summarize well the general content of the news.

Named Entity Recognition and normalization The standard classes of Named Entities (NE) (i.e. `person`, `location`, `organization` and `miscellaneous`) were the targets of chief actor extraction from news archive as well. Previous results proved that slight changes in the domain can lead to a significant drop in performance. Based on these issues, we decided to build subdomain specific corpora and employ different models in news channels. The selection of documents for manual NE annotation was done randomly from the archive, seeking a uniform spread in time (from 1999 to 2009) to avoid having too much news about one particular topic.

Conditional Random Fields [8], utilizing the rich feature set for Hungarian NE Recognition [18] were trained on the six subcorpora and on the whole corpus. NE extraction was carried out by different models according to their categories and the model trained on the whole NE corpus was applied to documents not belonging to the above-mentioned six categories (containing 23% of articles in total). In order to add NEs as tags to articles, their normalized forms had to be found. Two steps of normalization were performed, namely lemmatization and abbreviation resolution.

In morphologically rich languages such as Hungarian, nouns (including NEs) can have hundreds of different forms owing to grammatical number, possession marking and grammatical cases. When looking for the lemmas of NEs, the word form being investigated is deprived of all of the suffices it may bear. However, there are some NEs that end in an apparent suffix (such as '*McDonald's*' or '*Philips*' in English). The problem of proper name lemmatization is more complicated than that of common nouns, since NEs cannot be listed exhaustively, unlike common nouns, due to their diversity and steadily increasing number. NE lemmatization has not attracted much attention so far because it is not such a serious problem in major languages like English and Spanish as it is in agglutinative languages.

Our main hypothesis in NE lemmatization was that the lemma of an NE has a relatively high frequency in the whole news archive, compared to the frequency of the certain affixed forms of the NE. Hence, in order to be able to select the appropriate lemma for each NE phrase, we applied the following strategy: endings that seemed to be possible suffices were cut off from the NE; then the frequency of all possible lemmas in the news archive was counted and decision was made based on these frequencies, employing rules learnt from previous NE lemmatization experiments [3].

Lastly, abbreviation resolution was carried out in order to avoid duplicated entities in the global tag cloud (e.g. either '*United Nations*' or '*UN*' should be present in a tag set, but not both of them).

Extraction and derivation of noun phrases The tagging guide clearly states that tags have to be noun phrases (NP). Besides Named Entities, common nouns can be useful tags as topics of news articles can be described with their help. To extract NPs, we experimented with a Hungarian full constituent parser [1] and we tried to extract the deeper levels of the constituent tree. We found that not just the accuracy of that parser is average, but it is considerably slow as well. On the other hand, we observed that simple rules could gather the same or even better NPs. Hence, we simply extracted single nouns and successive adjective-noun and noun-noun phrases from articles. To obtain such morphological information, we applied a TNT-based POS tagger, trained on Szeged Treebank [7].

Apart from extracting NPs physically occurring in news items, we also derived NPs from verbs (from "*the bank was robbed*' to "*robbery*'), adjectives (from '*Italian*' to '*Italy*') and from other NPs (from '*the price of oil*' to '*oil price*'). The standard way of these kinds of transfers might be to stem the tag candidate in question, and match them with an element from a stemmed list of possible tags. In the case of matching, the original form of the term could be replaced by the stem found in the list. Unfortunately, the rich morphological nature of Hungarian language does not allow stemming and lemmas of different part-of-speeches are difficult to match. Instead, we applied here the most frequent hand-crafted transfer rules of derivation. A more sophisticated solution would be to invert an existing morphological analyzer [20] to get a full set of derivational rules.

Incorporating external knowledge Named Entity Recognizers and POS tagger-based heuristics are far from being perfect. However, gazetteers containing entities and topic identifiers do exist, e.g. they can be extracted from the Wikipedia. The full list of article titles of the Hungarian Wikipedia and the content of articles with title starting with "*List of*" (actually, in Hungarian ending with "*listája*") was cleaned and employed. Besides these lists, we used the set of users' tag from the manually labeled training set as gazetteer.

This kind of external knowledge was exploited by looking for items of the gazetteers in articles and we used their exact matches as potential tags. Of course, these matches can even introduce errors into the system as subphrases of a longer phrase may be present in the lists, while the longer one can be missing from them (e.g. '*New York*' and '*New York Times*').

Tag ranking We can extract potential tags from raw texts like those introduced above. However, the news articles are structured documents, i.e. they have title, heading, subtitles, they contains images with caption, links to other articles and formatting information (e.g. bold, italic), and so on. In order to exploit this structural information we investigated a weighting strategy. In this approach we assigned a weight for each formatting type and a relevance metric was calculated for each potential tag. We used this metric as a ranking function of tags. The parameterized tfidf metric of

$$tfidf(tag) = \frac{(\sum_{type} \lambda_{type} * tf(tag, type))^{\alpha}}{df(tag)^{\beta}}$$

was employed, where $tf(tag, type)$ refers to the frequency of a tag in the certain article in the given $type$ and $df(tag)$ is the number of documents that posses tag in their potential tag set, while α, β and λ are parameters to be optimized.

In order to find the optimal values for α, β and λ, we exploited the manually tagged corpus we had. We used those articles whose manual tag set was the real subset of the extracted potential tag set and treated the ranking of a potential tag set as `good` when the set of top ranked tags was equal to the manual tag set. The objective function of the parameter optimization problem was then the ratio of `good` rankings.

We found that only titles, heading, captions and italic regions should have a positive weight while the text itself and links just introduced noise into the system. This could be due to the issues that headings summarize well the content of articles and links may not target such closely related articles (e.g. they link recent news of the certain channels).

3.2 Wikipedia and Machine Learning-based abstract tagging

Appropriate tags often do not occur in the contents of a document to which it has been assigned. For example, an article dealing with the '*economic crisis*' may not contain the expression itself at all; however this tag would be the one which can cover the contents of that particular document in the best way. For this reason, extracting terms from the texts of documents is not enough. We call tags assigned to an article in such a manner that it is not contained in the document itself to *abstract tags* and the procedure of assigning this type of tags to documents was called *abstract tagging*.

3.2.1 Abstract tags based on Wikipedia

One of our modules responsible for assigning so-called abstract tags to news articles gets a set of potential tags extracted from articles themselves as input and derives a list of the titles of potentially useful, semantically related Wikipedia articles to them.

As a first step, the assignment of our candidate tags to Wikipedia articles was carried out. We mapped a candidate tag to a Wikipedia article if its normalized title matched our candidate tags. In those cases where we had an ambiguous tag candidate (i.e. having a disambiguation page on Wikipedia), we did not choose any of its Wikipedia article variants, in order to avoid involving noise in the later steps.

Then, five different abstract tagging methods based on the recognized Wikipedia articles were applied. The methods made use of the textual content of articles and the rich link structure existing among them as well. Following subsections discuss these heuristics.

Consideration of redirect pages Owing to the structure of Wikipedia, the very same contents might be obtained under different articles. For example, if we search for the term '*United States*' or '*Americans*', we get the same results. The pages responsible for redirection (redirect page) can be utilized to find synonyms (e.g. '*United States of America*' - '*United States*'), create associations (e.g. '*American*' - '*United States*'), resolve acronyms (e.g. '*USA*' - '*United States*') and to some content handle the misspelling of words (e.g. '*United states of America*' - '*United States*'). Based on these, we can determine a canonical representation of concepts, which has the benefit of increasing the cohesion of the whole tag set (e.g. '*gains*' and '*profit*' can have the same form). In our system, tag candidates whose corresponding Wikipedia article contained a redirection, the candidate terms were replaced with the title of the target of the redirection.

Extraction of definitions In the next phase, we extracted definitions for those tag candidates for which we had determined a Wikipedia article, and added them as abstract tag when more than one candidate tag of a document shared it. Extracting definitions can benefit in grabbing hyponym IS-A relations of concepts, e.g. it may be inferred that '*The Sopranos*' is an '*American TV-series*'.

Due to the encyclopedic nature of Wikipedia, there is usually a brief definition of the concept described in the actual document at the beginning of articles. In order to extract definitions, firstly we determined the sentence which was the most likely to contain valuable definitions. In our approach this sentence was the first one where the title of the article was presented, or when there was no such sentence, we chose the very first sentence from the first paragraph of the document. For instance:

Pál Erdős, one of the most outstanding mathematicians of the 20th century, member of MTA.

The Sopranos is an American TV-series, the creator and producer of which is David Chase.

Next, from the sentences selected in the above-mentioned way, we found all possible definitions. During this step, we used hyponym patterns (like '*is a*') as well as morphological and syntactic characteristics (e.g. the first noun occurring after the name of the Wikipedia article was mentioned in the content) of definition candidates. Another constraint was that all definitions gathered should be mapped into a Wikipedia article or each part of the potential definitions should have a Wikipedia article (e.g. '*American TV-series*' was considered as correct definition, since both '*American*' and '*TV-series*' had an article at Wikipedia.). For instance based on the two of examples above, the '*mathematics*' and '*American TV-series*', '*TV-series*', '*producer*' definitions were extracted, respectively.

Utilizing the link structure We also examined the possibility of assigning abstract tags by exploiting the rich link structure of Wikipedia. Here we employed three metrics:

1. we looked for those Wikipedia articles which frequently co-occurred with tag candidates in the form of links,
2. we examined those Wikipedia articles that were referred by more articles assigned to the set of tag candidates of a particular news document,
3. we also looked for articles that contained the most informatively a subset of potential tags of an article.

In the case of studying co-occurrences, we looked for Wikipedia articles that frequently co-occurred in the forms of links with some of our potential tags. This metric was utilized in such cases when a potential tag was referred to at least 10 times globally, but not more than 150 times. We did this because those articles that were referred less than 10 times seemed to be of low relevance, while those referred more than 150 times were too general.

For those articles that suited the limit according to its referrer pages, we looked for those distinct articles that were present in the form of a link at least half of the cases when the examined article was mentioned in form of a link. For example, since the co-occurrence measure for rally racer '*Sébastian Loeb*' and '*rally world championship*' was 0.7073, the latter term was also applied as an abstract tag for those news articles where '*Sébastian Loeb*' was extracted.

When examining outgoing links, we looked for articles that can be considered as relevant to a set of potential tags. We took every article referred to reliable outgoing links of the articles from the input set. We treated an outgoing link of an article as reliable if the article referred by it contained a back-reference to the referrer article, or if at least 25% of the links of a referring article pointed to the same article, and the number of the references was more than 3.

At the document level, an outgoing link was considered reliable and used its title as an abstract tag if more than one Wikipedia articles associated with potential tags of the news article contained it. For example in case of an article which contained both '*BUX*' and '*Stock Exchange of Budapest*', it induces the use of '*economy of Hungary*' as an abstract tag, since Wikipedia articles associated with both terms contain a reference to the same Wikipedia article.

As a third metric, we looked for Wikipedia articles (functioning as a potential abstract tag) with semantic relations to the set of input tag candidates by examining their outgoing links. In order to calculate a relatedness measure for an article, we used a modified version of the averages of the standard tf-idf metric:

$$tfidf'(d_j) = \frac{\sum_{t_i \in W} tfidf(t_i, d_j)}{|W|} \cdot |W \cap o(d_j)| \cdot \sum_{t_i \in W} lidf(t_i, 2),$$

where W refers to the set of Wikipedia articles (terms) assigned to tag candidates, $o(d)$ is the number of outgoing links of article d and $lidf(t, n)$ is the limited inverse document frequency of t, where limited means a constraint of n for the a minimum term frequency.

For each news article, the subset of those Wikipedia articles which contained at least one of the tag candidates in the form of an outgoing link was gathered. Lastly, the title of a Wikipedia article d_j was treated as abstract tag, if $tfidf'(d_j) > 0.3$.

3.2.2 Supervised learning of tags

Besides gathering abstract tags from Wikipedia, we gathered tags which represent the topic of the article (but do not occur in the document) by exploiting statistical patterns of the whole corpus as well. At the first glance, topic-related abstract tags should have been able to derive from the channel info of the articles, but in practice the channel hierarchy was intended for human browsing and not labeling of articles. There exists several diverse channels (with ten thousands of articles) and as the channel hierarchy has been evolving in the past 11 years, new channels spin out from parent channels, while others were deleted. For example, the category of `basketball` was introduced in 2001 and the related news from 1998 to 2001 were placed in `team sport` (that is its parent).

In order to extend the set of tag candidates with topic-related tags, we collected 243 tags to be learnt based on statistics. We defined a supervised machine learning task for each of the 243 tags. We trained classifiers to make the decision of adding a certain tag using the extracted set of potential tags as features. This kind of assignment has to be performed inside the particular top-level channel. For example, while the presence of the '*Manchester*' and '*Liverpool*' tag candidates may indicate tag '*Premier League*' in the `sport` channel, it does not do so in the case of political news. We employed articles having the tag in question as positive examples and articles at those days from other top-level channel as negative examples. Then the trained models (Logistic Regression) were used to make forecasts about each articles of the certain channel.

3.3 The final set of tags

After the extraction of potential tags from the document itself and the extension of this tag set by abstract tags, the average size of the tag sets per document was 17.3, while the tagging guide suggested using approximately 5 tags per article. Hence in the final step of the procedure the most reliable tags were selected.

The selection was carried out by hand-crafted heuristics based on the ranking of tags introduced in Section 3.1 (note that abstract tags cannot be ranked by the parameterized tf-idf metric). Besides the suggested number of tags per document, the selection had to fulfill other constraints such as at least one tag had to be placed in the `topic` category and tags with at least 3 of global frequency have to be used. The applied heuristics preferred to keep top-ranked NEs, which may came from the NE Recognizer and from automatically typed dictionary entries and abstract tags (see the most frequent post-processing method in Section 3.1.1), in `person`, `organization`, and `location` tag types. For the `topic` tag type, the selection was carried out among the top-ranked common nouns, frequent abstract tags and top-ranked miscellaneous NEs. After the selection, the average size of the final tag sets became 5.2.

4 EVALUATION

The manual and automatic tagging of the articles at [origo] can be characterized by the basic statistics presented in Table 1. These figures show that automatic tagging has a very similar nature to the manual one. The only exception is with the average number of daily created new tags. This is due to the fact that the set of tags used attained an appropriate state, so new tags were rarely introduced.

Table 1. Statistics of manual and automatic tagging

	manual	automatic
Start date	15-2-2009	5-12-1998
End date	22-10-2009	14-2-2009
Number of articles tagged	28,055	366,937
Number of tags created	15,726	66,843
Number of new articles created (daily average)	110	93
Number of new tags created (daily average)	45	17
Average number of tags Assigned to an article	3.42	4.98
Average length of tags used (tokens)	1.48	1.45
Distribution of tag types of manual tagging TOPIC	8495 (54%)	34276 (51%)
PER	3637 (23%)	16933 (25%)
ORG	2281 (15%)	11549 (17%)
LOC	1313 (8%)	4085 (7%)

4.1 Evaluation of automatic tagging

Quantitative evaluation of automatic free-text-tagging cannot be carried out using automatic metrics, because the comparison of manually and machine assigned tags requires the recognition of synonyms and hypernyms. More importantly, the judgment of the relevance and necessity of a tag is definitely subjective.

For comparability considerations, we decided to adopt the state-of-the-art keyphrase extracting system, KEA[23] for our task (i.e. we re-trained KEA on the labeled set and applied Hungarian language-dependent features). In order to check the reliability of the two automatic tagging procedures, the authors of the tagging guideline of the Origo Ltd. manually checked the tagging of 725 randomly selected articles. KEA achieved a tag-level F-measure of 32.0 (precision of 22.6% and recall of 54.9%) while our system achieved an F-measure of 71.9 (precision of 59.4% and recall of 75.4%). If the mismatches of the tag's type were also taken into account, the our results drop to 66.2 (KEA is not able to distinguish types).

4.2 Evaluation of Wikipedia-based abstract tagging

We consider the investigation of Wikipedia-based abstract tagging as novel results, hence we carried out a quantitative evaluation on this submodule as well. For the difficulties of evaluating abstract tagging, the same holds as for the evaluation of the whole tagging procedure itself. For this reason, due to the especially high subjective nature of the evaluation of abstract tagging, two linguists were asked to decide on the appropriateness of each tag assigned to news articles by the abstract tagging module relying sorely on Wikipedia articles. 600-600 documents were chosen for evaluation, out of which 100 were the same for both annotators. This way we would have 1100 different documents for evaluation. There were 1114 abstract tags (tag which does not present in the text of the article) among the manually assigned tags on this documentset. During the abstract tagging procedure enhanced by Wikipedia, there were all together 5014 assignment of 2028 distinct abstract tags in case of the test set.

The procedure of evaluation was as follows: annotators had to examine each abstract tag assigned to an article, and decide, whether it was an acceptable tag with respect to the content of the document (precision), taking the tagging guidelines of [origo] into consideration as well. Simultaneously, they had to decide if the automatic abstract tags of a document were able to cover the meaning exactly or partially one or more abstract tags assigned manually by editors at [origo] (recall). The different metrics used for Wikipedia-based abstract tagging can be just evaluated by precision (see Table 2).

Table 2. Results achieved by different abstract tagging heuristics

Heuristic	#tag	precision
Redirection	1155	72.38
Definition	1471	28.14
Co-occurrence	1998	34.88
Outgoing links	558	40.68
Container	551	16.33

The final measure of the quality of abstract tagging was computed by F-measure, combining the precision of automatic abstract tags and the extent to which abstract tags were able to cover manual abstract tags. Finally, an F-measure of 16.75 was achieved.

These results are satisfactory, if we take the fact into consideration that coverage was compared to manual tagging of [origo] employees, who has of course access (as human beings) to a full sense repository and not just Wikipedia (only the 20.76% of the manually assigned abstract tags had a corresponding Wikipedia article).

5 UNSUCCESSFUL METHODS

In addition to the procedures introduced in Section 4, we experimented with several other methods. Improvements were expected from these procedures, but we found them unsuccessful (at least for this certain dataset). Here, we briefly describe three of them (supervised learning of tag relevance, employing existing ontologies, analysis of the link graph), as they may be interesting negative results.

There are several nouns which can never be used as a tag (e.g. '*news*'). On the other hand, from the manually labeled documentset, we had 2828 tags which were used in more than 2 documents. We defined a supervised machine learning setting in order to learn the relevance of a noun, i.e. whether it can be used as tag. We used the tags of manual annotation as the positive sample and manually chose 300 negative examples. We constructed a rich feature set consisting of corpus frequency-based measures like the distribution in time, distribution among categories, average term frequency and document frequency. In the evaluation (ten-fold-cross-validation), the system could not outperform the most-frequent-class baseline. Thus our hypothesis that "tag relevance can be learnt from frequency patterns" may be wrong and it can be learnt just on the basis of semantic analysis.

Another idea was to use existing taxonomies (like the Hungarian Thesaurus [22]) and ontologies (the Hungarian WordNet [12]) as the basis for a calculation of semantic relatedness among tag candidates. However the recall of these resources is quite low and the WordNet uses an over refined sense set, thus even it contains the phrase in question, the correct synset is difficult to select. Instead of using these resources, we decided to exploit the link structure of Wikipedia to derive semantic relatedness among phrases.

Lastly, we expected to get useful information from the analysis of the link graph of the news archive. Our hypothesis here was, that articles belonging to one particular narrow topic are strongly inter-linked (for example storytelling). However, the main goal of a news portal is to keep the user surfing the site for as long as possible. For this, articles which will be probably read by the user should be offered, but such offers (links) do not indicate a topic similarity among articles as users like to jump from topic to topic (e.g. they read the hot topics of a day in order). Moreover, the level of linkage was growing during the life of the news portal, thus its characteristics were very different in 1998 and in 2009. The width of the link graph (the maximum of shortest paths among pairs of articles) is 7. This means that the article of a *basketball match summary* is accessible from the article about the *previous day of the Hungarian Parliament* in less than 7 clicks (here the link between sport and politics is the news article about a *friendly soccer match between political parties*).

6 CONCLUSIONS

Here we introduced the the task of automatic free-text-tagging of the news archives. From an application viewpoint, the tagging of news has several benefits, such as that for contextual advertising, the organization of the news set, behavioral targeting and increasing connectivity. From a research point of view, it differs from the tasks of tag recommendation and key-phrase extraction, and it has several special characteristics (the importance of NEs, structured documents, etc.).

The 370 thousands of articles in the news archive could not be tagged by the community of readers or by a team of journalists. We showed that the free-text-tagging could be carried out by an automatic system and we achieved a satisfactory F-measure of 71.9. This result is revalued if we take into account the fact that Hungarian - the language of the news archive - has special characteristics (e.g. agglutinivity and free word order) and the set of available language processing tools and related resources are limited (e.g. the size of Hungarian Wikipedia is 4% of the English one).

Our system consists of several modules which solve particular subtasks. Among these solutions, we consider the abstract tagging exploiting Wikipedia and tag ranking as remarkable results. Still, there are several straightforward ways in which our system could be improved. For example, we plan to perform the final selection of tags among the potential tags in a more sophisticated way exploiting semantic relatedness – calculated from the corpus and from Wikipedia – among tags. Another useful task would be to incorporate inter-document information (document similarity) into the system as corpus level issues are currently exploited just in the supervised learning of tags. Lastly, we plan to investigate the potential utility of using the context of tag candidates in their ranking.

7 ACKNOWLEDGMENTS

This work was supported in part by an NKTH grant (project codename TEXTREND) of the Hungarian government. The authors also wish to thank Veronika Vincze and Attila Almási (who did the NE annotation work) for their devoted efforts.

REFERENCES

[1] Anna Babarczy, Bálint Gábor, Gábor Hamp, and András Rung, 'Hunpars: a rule-based sentence parser for hungarian.', in *Proceedings of the 6th International Symposium on Computational Intelligence*, (2005).

[2] Folke Eisterlehner, Andreas Hotho, and Robert Jschke, eds. *ECML PKDD Discovery Challenge 2009 (DC09)*, volume 497 of *CEUR-WS.org*, September 2009.

[3] Richárd Farkas, Veronika Vincze, István Nagy, Róbert Ormándi, György Szarvas, and Attila Almási, 'Web based lemmatisation of named entities.', in *Proceedings of the 11th International Conference on Text, Speech and Dialogue*, pp. 53–60, (2008).

[4] Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-manning, 'Domain-specific keyphrase extraction', in *Proceeding of 16th International Joint Conference on Artificial Intelligence*, pp. 668–673. Morgan Kaufmann Publishers, (1999).

[5] Maria Grineva, Maxim Grinev, and Dmitry Lizorkin, 'Extracting key terms from noisy and multi-theme documents', in *18th International World Wide Web Conference (WWW2009)*, (April 2009).

[6] Margaret Kipp, '@toread and cool: Tagging for time, task and emotion', in *Proceedings 8th Information Architecture Summit*, (2007).

[7] András Kuba, András Hócza, and János Csirik, 'Pos tagging of hungarian with combined statistical and rule-based methods', in *Proceedings of the 7th International Conference on Text, Speech and Dialogue*, pp. 113–120, (2004).

[8] John Lafferty, Andrew McCallum, and Fernando Pereira, 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data', in *Proc. 18th International Conf. on Machine Learning*, pp. 282–289. Morgan Kaufmann, San Francisco, CA, (2001).

[9] T. K. Landauer and S. T. Dumais, 'Solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge', *Psychological Review*, (1997).

[10] Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun, 'Clustering to find exemplar terms for keyphrase extraction', in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 257–266, Singapore, (August 2009).

[11] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis, 'Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead', in *Collaborative Web Tagging Workshop at WWW2006*, (May 2006).

[12] M. Miháltz, Cs. Hatvani, J. Kuti, Gy. Szarvas, J. Csirik, G. Prószeky, and T. Váradi, 'Methods and results of the hungarian wordnet project', in *Proceedings of the 4th Global Wordnet Conference*, (2008).

[13] Peter Mika, 'Ontologies are us: A unified model of social networks and semantics', *Web Semantics: Science, Services and Agents on the World Wide Web*, **5**(1), 5–15, (2007).

[14] Gilad Mishne, 'Autotag: a collaborative approach to automated tag assignment for weblog posts', in *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pp. 953–954, New York, NY, USA, (2006). ACM Press.

[15] Philip Resnik, 'Using information content to evaluate semantic similarity in a taxonomy', in *IJCAI*, pp. 448–453, (1995).

[16] Sanjay Sood, Sara Owsley, Kristian Hammond, and Larry Birnbaum, 'Tagassist: Automatic tag suggestion for blog posts', in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, (2007).

[17] Michael Strube and Simone Paolo Ponzetto, 'Wikirelate! computing semantic relatedness using wikipedia', in *AAAI'06: proceedings of the 21st national conference on Artificial intelligence*, pp. 1419–1424. AAAI Press, (2006).

[18] György Szarvas, Richárd Farkas, and András Kocsor, 'A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms', *DS2006, LNAI*, **4265**, 267–278, (2006).

[19] M. Tatu, M. Srikanth, and T. D'Silva, 'Rsdc'08: Tag recommendations using bookmark content', in *Proceedings of the ECML PKDD Discovery Challenge 2008*, (2008).

[20] Viktor Trón, Laszló Németh, Péter Halácsy, András Kornai, György Gyepesi, and Daniel Varga, 'Hunmorph: open source word analysis.', in *Proceeding of Association for Computational Linguistics*, (2005).

[21] Peter Turney, 'Coherent keyphrase extraction via web mining', in *Proceedings of IJCAI*, pp. 434–439, (2003).

[22] Rudolf Ungváry and Tamás Radnai, 'Discover thesauri: State of the art in hungary', in *Proceedings of 3rd Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence*, (2005).

[23] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig, 'Kea: Practical automatic keyphrase extraction', in *ACM DL*, pp. 254–255, (1999).