Expanding the horizons: adding a new language to the news personalization system

Andrew Fedorovsky
News360 Ltd.
afedorovsky@news360.com

Maxim Ionov
News360 Ltd.
max.ionov@gmail.com

Varvara Litvinova
News360 Ltd.
vlitvinova@news360.ru

Tatyana Olenina
News360 Ltd.
tolenina@news360.ru

Darya Trofimova
News360 Ltd.
dtrofimova@news360.ru

Abstract

News360 is the news aggregation system with personalization. Initially created for English, it was recently adapted for German. In this paper, we show that it is possible to adapt such systems automatically, without any manual labour, using only open knowledge bases and Wikipedia dumps. We propose a method for adaptation named entity linking and classification to target language. We show that even though the quality of German system is worse than the quality of English one, this method allows to bootstrap a new language for the system very quickly and fully automatically.

1 Introduction

Every day news sources generates millions of news articles. News aggregation systems helps users to examine this overwhelming amount of information, combining thousands of article feeds into one feed of news events. The next evolutional stage of such systems are personalized news aggregators, which forms overall news feed based on users preferences.

News360 was created as one of these personalized news aggregation systems. Our crawler collects articles from tens of thousands of news sources, join them into clusters associated with news events and present them to user, ranking in order of her preferences. A brief description of modules of the system will be given in the section 1.1.

We have started working with English news articles and spent a lot of time improving our classification, clustering and personalization quality for users in USA, UK and other English-speaking countries. However, to further increase number of our users we had to add another language into system. So the problem was how to make our system multilingual and reach quality level for the new languages comparable with quality, that was already reached for English news. The approach proposed in this paper is fully automatic. Using it, we have successfully built German version of our system, which is already available for our users in Germany. Our approach allows us to easily add other languages and we expect that in a nearest year we will be able to work with 3-4 more European languages and probably one Asian. Before going into the details of the approach itself, we should describe our news article processing pipeline.

1.1 News360 Overview

News360 pipeline consists of 5 stages:

- **Crawling** articles from news sources, parsing them for text, attributes and metadata;
- **Named-entity linking (NEL)**;
- **Classification** and tagging news articles;
- **clustering**: group articles about same news event into one cluster;
- **Personalization**: retrieve results to users request, ranking them by a bunch of parameters, including users preferences.

We will not describe crawling, clustering and personalization stages here, because we assume them language independent (see section 1.4).

1.2 Named Entity Linking (NEL)

Named Entity Linking is the task of linking entities from some knowledge base to their mentions in the text. A lot of work in this field was done using open knowledge bases like DBPedia, Freebase

Proceedings of the First Workshop on Computing News Storylines, pages 35–39,
Beijing, China, July 31, 2015.

or Wikipedia (see, for example, (Shen et al., 2015) for a survey).

NEL component in our system links mentions to entities in the manually curated ontology that was partly extracted from Freebase[1] and Crunchbase[2]. We have extracted only named entities: persons, locations, products and organizations. All mentions for an entity that were either extracted from an ontology or added manually are stored in the ontology as "synonyms" for an entity. During the processing of a news article, the system finds all the possible synonyms for all the entities in text. After that, all found objects are ranked by a set of hand-crafted rules. The structure and the evaluation of these ranking rules are out of the scope of this paper as we have turned off all rules that could be language dependent. Another component that we will not discuss here is the component that identifies unknown objects. Since it is rule-based and designed for English, it was useless in the multilingual scenario.

1.3 Classification

Apart from ontology, there is a wide tree of categories in our system. Total number is over 1000, and this number is increasing constantly. It includes both wide topics like "Space" and "Tech" and very marginal topics like "Microwaves" and "Blenders".

There are different modules that detect categories for an article in our system, each can add or remove[3] one or more category. The one that was most important for English articles was based on hand-crafted keywords, which, as we thought, could not be ported to other language fast. Another system was based on objects. It used automatically obtained mappings from objects to categories. We have set our hopes on this system because of its complete language independence.

1.4 Language (In)dependence

We have assumed that the only language dependent components of the system are linguistic components: NEL and classification, whereas other parts of the process, for example, personalization and clustering are language independent. This may be an oversimplification, because it is possible that language influences user preferences and expectations[4]. Still, we think that this question is not of paramount importance. We discuss this briefly in section 4.

Given this, the process of adding new language limits to this surprisingly small amount of steps:

- Implement Named Entities Linking to the objects in the ontology for the new language
- Implement classification process based on keywords to classify news articles in the new language

In the next sections we show that these two processes are sufficient to include news in any language to our pipeline. Section 2 is devoted to the problems we faced and decisions we made to overcome them. In section 3 we evaluate the system. In Section 4 we present our conclusions and discuss possible future improvements.

2 Methods of Extracting German Data

We have decided to employ the existing ontology for German instead of creating a new, unified ontology for both languages from scratch. For years of work, the ontology that we used was fine-tuned and upgraded, dumping all these changes would be unwise and would create a lot of bugs in the system.

2.1 Extracting German Data: Entity Linking

As it was already stated, to extract objects from English texts, our NEL component looks for every possible mention of any object in the ontology. These mentions are the "aliases" for entities, or "synonyms" as we call them. Since we have decided to use the ontology built for English articles, the only missing component were the synonyms for the target language. In order to extract them, we used several sources: Wikipedia dump, Freebase (Bollacker et al., 2008) and Wikidata (Vrandečić and Krötzsch, 2014) ontologies[5].

Since most of objects in our ontology were initially extracted from Freebase, links to the original Freebase entities were already known. Some of these objects in Freebase link to Wikipedia. On every step we have lost some fraction of objects: some objects in our ontology did not have a link to

[1] https://www.freebase.com/

[2] https://www.crunchbase.com/

[3] This is helpful sometimes to avoid presence of two controversial categories

[4] cf. Sapir-Wharf hypothesis of language relativity (Whorf and Carroll, 1956)

[5] When we started this project we have not know yet that Freebase was going to be discontinued. After the announcement, we added Wikidata to the list of our sources. We could not switch to it entirely since it had less data than Freebase

Freebase, some links has changed since the extraction, etc. Number of mapped entities, compared to the total number of entities in the ontology is presented in the table 1.

We also tried to map entities from our ontology with Wikipedia articles simply by their names and aliases, but mapping only by name showed low precision whereas mapping by aliases showed very low recall.

After establishing links from our ontology to Wikipedia articles we were able to extract possible object names from two different sources:

- Aliases for the object in Wikidata,
- Redirects to the object page in Wikipedia in target language.

Aliases were obtained by parsing JSON dump of Wikidata, the list of redirects were extracted with wikipedia-redirect utility[6]. Number of extracted synonyms are presented in the table 1.

Stage	$N_{entities}$	$N_{synonyms}$
English	662,462	5,008,436
German	111,126	278,964

Table 1: Amount of synonyms.

2.2 Extracting German Data: Classification

As it was said before, classification system based on hand-crafted keywords for every category was the most important. There were two ways of getting this system to work in German:

1. Porting existing keywords to another language;

2. Extracting keywords for another language automatically.

To port existing keywords we have decided to translate them automatically, using Yandex translation services[7]. Understanding that the translation would not be perfect, we have assumed that this is the most rapid way to approach an acceptable rate of classification quality for the new language. To further improve classification quality, we have tried to extract new keywords automatically. This process is described in the next section.

[6] https://code.google.com/p/wikipedia-redirect/

[7] https://tech.yandex.com/translate/

2.3 Various Sources of New Data

To extract keywords in the desired language automatically, we used Wikipedia as a corpus tagged by categories. Using Wikipedia categories as an approximate thematic markup, we mapped 80% of our topics to one or more Wikipedia categories. This way for every mapped category we have acquired a corpus which could be used to extract keywords. The topic was considered to be mapped on a Wikipedia category if the category contained the stem of the topic name as a substring.

After that one should determine keywords. We did not solve this task for topics which contained too little data from Wikipedia these texts, were used as a background corpus together with texts from topics which could not be mapped. We also ignored infrequent words. The first metric we used to score word relevance to topic was TF-IDF of given word in given topic, the second one was the conditional probability for text to be in the topic given that text contains the word.

As our most important categorization system is case sensitive, it is reasonable to take capitalization into consideration, especially for German. However, there is a risk to lose the word if it is specific for the topic but occurs in different capitalizations so as none of them look very important. Thus we counted TF-IDF for every form of every word and the second metric for lowercased forms. Words with the highest TF-IDF were marked as the keywords if their lowercase forms had high rank in the second list.

All these keywords got a moderate positive weight and gave the increase of categorization recall with no precision decrease, which caused a gain in F0.5-score for about 3%. Lowering a minimum threshold for word to be taken as a keyword gave nothing at first as they began to intersect with already existing sets but then gave drastic decrease of precision. See table 3 for details.

Using topic-specific N-grams should increase an impact of this method on the overall quality.

3 Experiments and Evaluation

3.1 Creating Evaluation Corpora

To evaluate the performance of the system on the new language, we had to evaluate system performance on English news articles first, since it was not done before. To do this, we have collected and marked up corpora. Our English corpus consisted of 100 non-random news articles, covering

most basic categories: politics, sports, business, tech and so on. German corpus was smaller, it consisted of 24 non-random articles. Its size influenced its coverage: some important topics were not represented in the corpus at all.

Each article in each corpus was processed with the system and then fixed by hand by two experts independently. All inter-annotator disagreements in markup were settled. The procedure of corpora markup may have influenced the result: errors and focus of the system may have influenced the opinion of experts, but we will assume that possible error is insignificant, leaving this question for further research. Entities were marked up and linked to the ontology in each article, all possible topics were found for each article.

We have computed standard metrics for evaluation: precision and recall, but instead of using F1-measure as an average, we have chosen F0.5-measure (Rijsbergen, 1979). Precision is more important for the system than recall: showing something wrong to the user is much worse than not showing something.

Also, apart from measuring performance of the system on English and German, we measured it with so called "Emulated conditions": a system working with English while everything non-reproducible in German (or any other target language) was disabled. For example, the entity in the ontology was available in this setup only if it have been interlinked with an entity in target language (so we could extract synonyms for it). Using these conditions we could get approximate evaluation without corpora on the target language.

3.2 Named Entity Linking Task

The NEL component for German articles shows quality comparable to English given that there are six times less entities in German than in English in the ontology (as seen in table 1). The results for different setups are given in the table 2. Text was treated as a bag of non-unique objects: score for each object in corpus was the number of times the object was found in text divided by the number of object in corpus.

Experiment	P	R	F0.5
English	0.938	0.662	0.866
Emulated conditions	0.849	0.607	0.786
German	0.790	0.422	0.673

Table 2: NEL evaluation.

3.3 Classification Task

Classification performs much worse than NEL (see table 3). Experiments (2) and (3) used language-independent classification components only, first of all categorization based solely on objects detected in texts. This method showed poor results probably because of types of objects in our ontology: they are all named entities, but not every category has a lot of named entities connected to it. Different categories vary in the average number of objects in texts, so this method works well only for a limited number of categories.

Categorization based on keywords, in contrast to the object-based method, behave quite unexpectedly: even when used with English keywords, it increases the quality of categorization drastically (4). Using keywords translated with machine translation increases the quality further (5). Methods described in section 2.3 allow to increase the quality further (6).

Experiment	P	R	F0.5
(1) English	0.766	0.619	0.731
(2) Emulated conditions	0.545	0.189	0.396
(3) German, no keywords	0.429	0.058	0.188
(4) German, English kw	0.483	0.182	0.363
(5) German, translated kw	0.569	0.240	0.447
(6) the same + new kw	0.562	0.325	0.490

Table 3: Classification evaluation.

4 Conclusion and Future Work

We have showed that new languages can be integrated without great effort into systems similar to ours. Both NEL and classification modules show acceptable quality that are sufficient for launch.

Another result of this paper is the demonstration of applicability of machine translation to such unexpected tasks as providing keywords for classification.

One interesting topic that was left for further research is how appropriate it is to use the same ontology for different languages. It is possible that native speakers of two different languages would require two slightly different ontologies because of different way of thinking. Still, this approach is worse from engineering point of view: not only this is an unnecessary redundancy, this is also the possible source of undesired divergences in ontologies. So, despite the possible theoretical problem, having shared ontology seems more practical.

References

Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In Jason Tsong-Li Wang, editor, *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

C. J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460, Feb.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September.

B.L. Whorf and J.B. Carroll. 1956. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Language/Anthropology. Technology Press of Massachusetts Institute of Technology.