Found in Translation

Marco Turchi¹, Ilias Flaounas², Omar Ali¹, Tijl De Bie¹,
Tristan Snowsill¹, and Nello Cristianini^{1,2}

¹ Department of Engineering Mathematics, Queen's Building

² Department of Computer Science, Merchant Venturers Building
Bristol University, Bristol, United Kingdom
<http://patterns.enm.bris.ac.uk>

Abstract. We present a complete working system that gathers multilingual news items from the Web, translates them into English, categorises them by topic and geographic location and presents them to the final user in a uniform way. Currently, the system crawls 560 news outlets, in 22 different languages, from the 27 European Union countries. Data gathering is based on RSS crawlers, machine translation on Moses and the text categorisation on SVMs. The system also presents on a European map statistical information about the amount of attention devoted to the various topics in each of the 27 EU countries. The integration of Support Vector Machines, Statistical Machine Translation, Web Technologies and Computer Graphics delivers a complete system where modern Statistical Machine Learning is used at multiple levels and is a crucial enabling part of the resulting functionality.

1 Introduction

This paper describes a web system that makes use of Statistical Machine Learning technology at various levels, in order to process and aggregate the contents of hundreds of news outlets in dozens of languages: Found in Translation (FIT) (<http://foundintranslation.enm.bris.ac.uk/>).

News outlets produce a huge amount of daily news in different languages which are usually read only by the citizens of the country where the media that publish the story is located. Found in Translation is an online service that makes all this information available to English speakers under a common interface.

Many systems have been proposed to show different aspects of the information contained in various news outlets. NewsBlaster [1] generates multi-document summaries, while European Media Monitor [2] uses cross-language retrieval to generate news clusters without translating. The TextMap system [3] is mostly based on detecting entities.

Found in Translation takes advantage of the technologies obtained by the Statistical Machine Translation (SMT) and text categorisation communities in recent years. All of these advances have been possible only due to statistical machine learning technology. Instead of using cross-language information retrieval techniques to analyse multilingual news, FIT translates daily news from 21 European languages into English. A set of classifiers categorises each translated article

into six topics, namely ‘Accidents’, ‘Business’, ‘Crime’, ‘Environmental Issues’, ‘Politics’ and ‘War and Conflict’. All the news are geolocated and presented according to their topic on a ‘heat-map’ of Europe to allow an easy comparison of the topic bias present in the media in the different countries of EU.

Our approach has the advantage that an English speaker can easily read content translated into English from the different languages spoken in the EU. He or she can also easily compare the current levels of interest in the media for a specific topic, in each of the EU countries. Comparisons like this could be useful in understanding the different perspectives and biases that the media present to each country, and such an analysis could not be performed by a single human being.

From the point of view of machine learning, the problem of creating a classifier for each topic and for each language is very difficult to tackle: It is almost infeasible to gather all the multilingual labelled data that are needed to train the classifiers. The use of machine translated news allows us to focus on only one language, English, where labelled data are more easily available.

2 Methodology

FIT is based on state of the art technology both in text categorisation, using Support Vector Machines (SVMs) [4] and in SMT, using Moses [5].

We consider the popular class of SMT approaches based on a noisy channel model originally developed for speech recognition [6], where a Markovian Language Model coupled with a phrase-to-phrase translation table are at the heart. In recent years, the noisy channel model has been extended in different directions. The most fruitful has been the phrase based statistical machine translation (PBSMT) introduced by [7] that is based on the use of phrases rather than words. We use Moses, a complete phrase based translation toolkit for academic purposes. It provides all the state of the art components needed to create a PBSMT system from one language to another. For each language pair, an instance of Moses is trained using Europarl [8] data and JRC-Acquis Multilingual Parallel Corpus [9]. To train the language model, this data is complemented with the English news of the last fifteen days.

We use a set of Support Vector Machines (SVMs) classifiers, one per topic we want to detect. SVMs are kernel based approaches of supervised learning, delivering state of the art performance in text-categorisation [4]. Each classifier is trained offline, on a dataset of articles which have been pre-labelled. Linear kernels were used and the classifiers were trained with a choice of parameters aimed at achieving a low false positive rate, in other words a high precision of above 90%. This means that documents will be tagged only when we are very confident about the topic, while articles for which the confidence is lower will be left untagged.

To calculate an unbiased influence of a topic in a country, a subset of news outlets is selected. For each country, all the outlets of that country are ordered using their Alexa rank score (<http://www.alexa.com>). This score, which has been used before in the spatial analysis of news sources [3], is calculated using

a combination of criteria and it estimates the traffic rank of each website. For FIT, we use as input dataset only the top ten ranked outlets in that country. For each country and each topic, we count the number of articles which referred to that topic during a sliding window, and we normalize it by the total number of articles discovered in that country in these ten outlets. This quantity is used on the web interface to colour the European map.

3 Infrastructure

Nowadays most main news outlets provide their content in Really Simple Syndication (RSS) format. We developed a web crawler infrastructure that targets these RSS feeds and collects all the advertised articles on the main page of each outlet we track. The RSS feeds provide just the summary of each article and a link to the web page of the full article. The crawler uses that link to get the full HTML page that contains the main body of the article. Finally a scraper is used to remove undesired elements from the HTML page such as images, advertisements, external links, etc., and returns the raw text of the article.

Each article we collect is stored in a database for further processing. The information stored includes a title and a summary (from the RSS feed), the full text of the article (from scraping the relevant HTML page), the publication date, the outlet name of its origin, the geographic location of the outlet and the language it is written in.

Every day, roughly 10,000 multilingual news articles are extracted and translated into English before stored in the database. Subsequently, all the English and machine-translated English articles are tagged by topic using the six SVM classifiers described previously.

4 Web Interface

FIT has a web-based user interface that presents all the translated news articles in an organised way. Articles are divided by topic and each topic is further divided by country. Two different data representations are available. The first is using colour maps of Europe, indicating topic activity. Each topic is shown on a different map as for example in Fig. 1. The second representation visualizes the same information using bar charts (see web site).

On the map, each country is coloured according to an unbiased statistic computed only on the top-ten outlets of the country as ranked by their Alexa score. There are six different maps, one map per topic. Green colours identify a weak interest for a particular topic while red colours a strong interest. On the bar chart view, the size of each bar represents the total number of discovered, translated and categorised news articles for each country for a particular topic. The list of articles we show for each country and topic is produced by using only high quality machine translated news: A filter has been introduced to hide articles that contain more than one untranslated word.

Fig. 1. Found in Translation web site

For each article the translated title and summary are reported with links to their site of origin. For United Kingdom and Ireland only English language news are shown. Multi-lingual countries, like Belgium, contain news from more than one language, all translated in English. This aspect is completely transparent to the user.

Acknowledgements. We are grateful to Simon Price and IRLT for their contribution to the web interface. This work is supported by the EU projects SMART and PASCAL2. Flaounas is supported by the A. S. Onassis Public Benefit Foundation. Cristianini is supported by a Royal Society Wolfson Merit Award.

References

1. McKeown, K., Barzilay, R., et al.: Tracking and summarizing news on a daily basis with columbia's newblaster. In: Proceedings of HLT 2002, San Diego, USA (2002)
2. Best, C., der Goot, E.V., et al.: Europe media monitor. Technical report, JRC (2002)
3. Mehler, A., Bao, Y., et al.: Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics* 12(5), 765–772 (2006)
4. Joachims, T.: SVM light (2002), <http://svmlight.joachims.org>
5. Koehn, P., Hoang, H., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings ACL 2007, demonstration session (2007)
6. Brown, P.F., Pietra, S.D., et al.: The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2), 263–311 (1994)
7. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings NAACL 2003, Morristown, NJ, USA, pp. 48–54 (2003)
8. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Machine Translation Summit X, pp. 79–86 (2005)
9. Steinberger, R., Poulighen, B., et al.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Arxiv preprint cs/0609058 (2006)