University of BRISTOL

Found In Translation

ECML-PKDD 2009

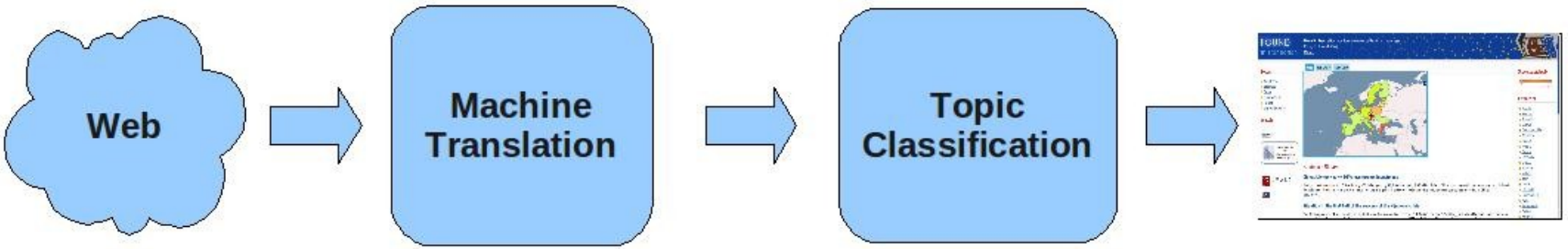
Tuesday, September 8

M. Turchi, I. Flaounas, O. Ali,

T. De Bie, T. Snowsill, N. Cristianini

Found In Translation

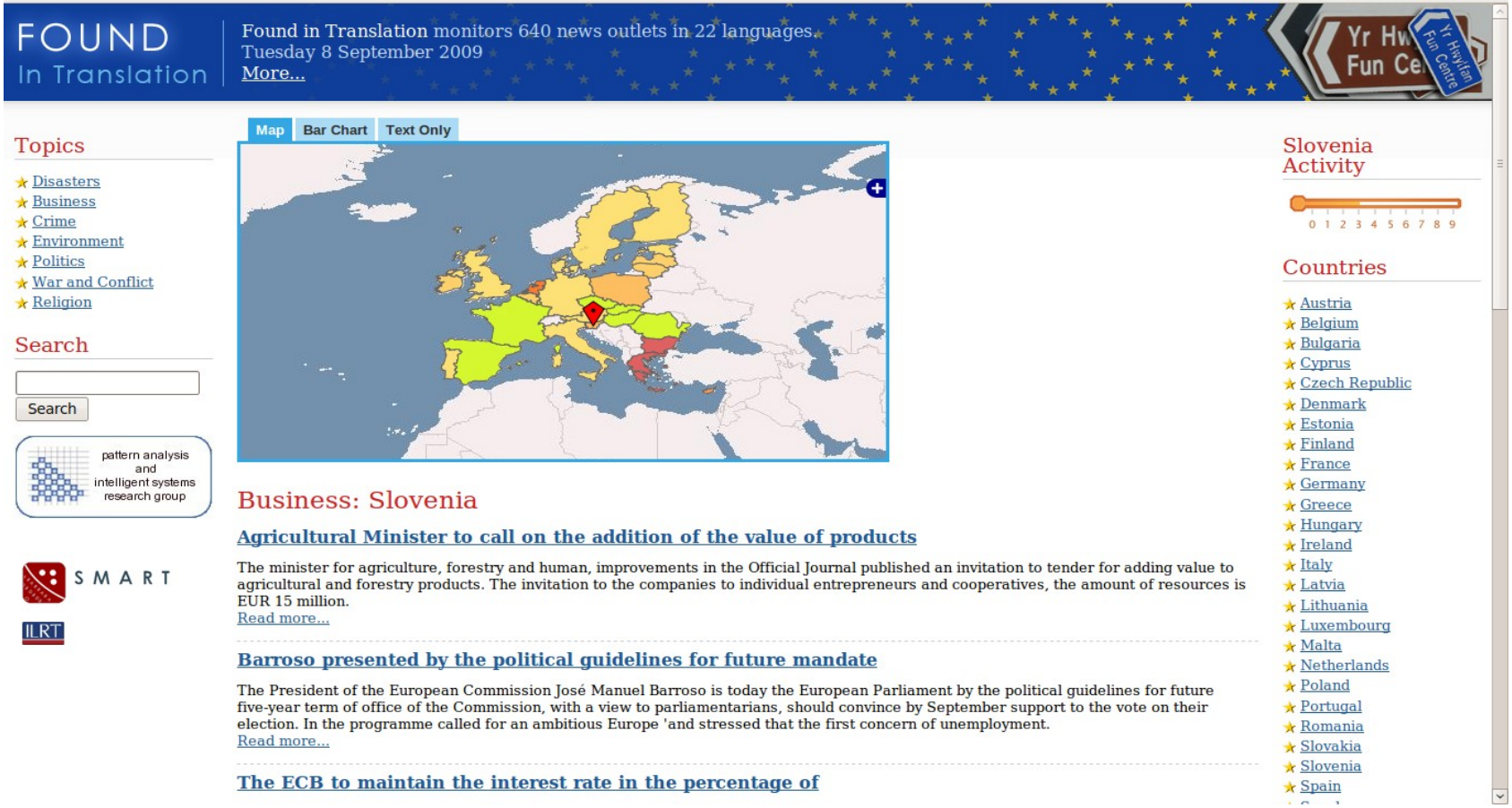
FIT is an example of integration of state of the art technologies in machine intelligence, that enables a different user experience of the mediasphere.



We monitor 640 news outlets in 22 languages.
Open Source Software was used: Moses and SVM-Light

Found In Translation

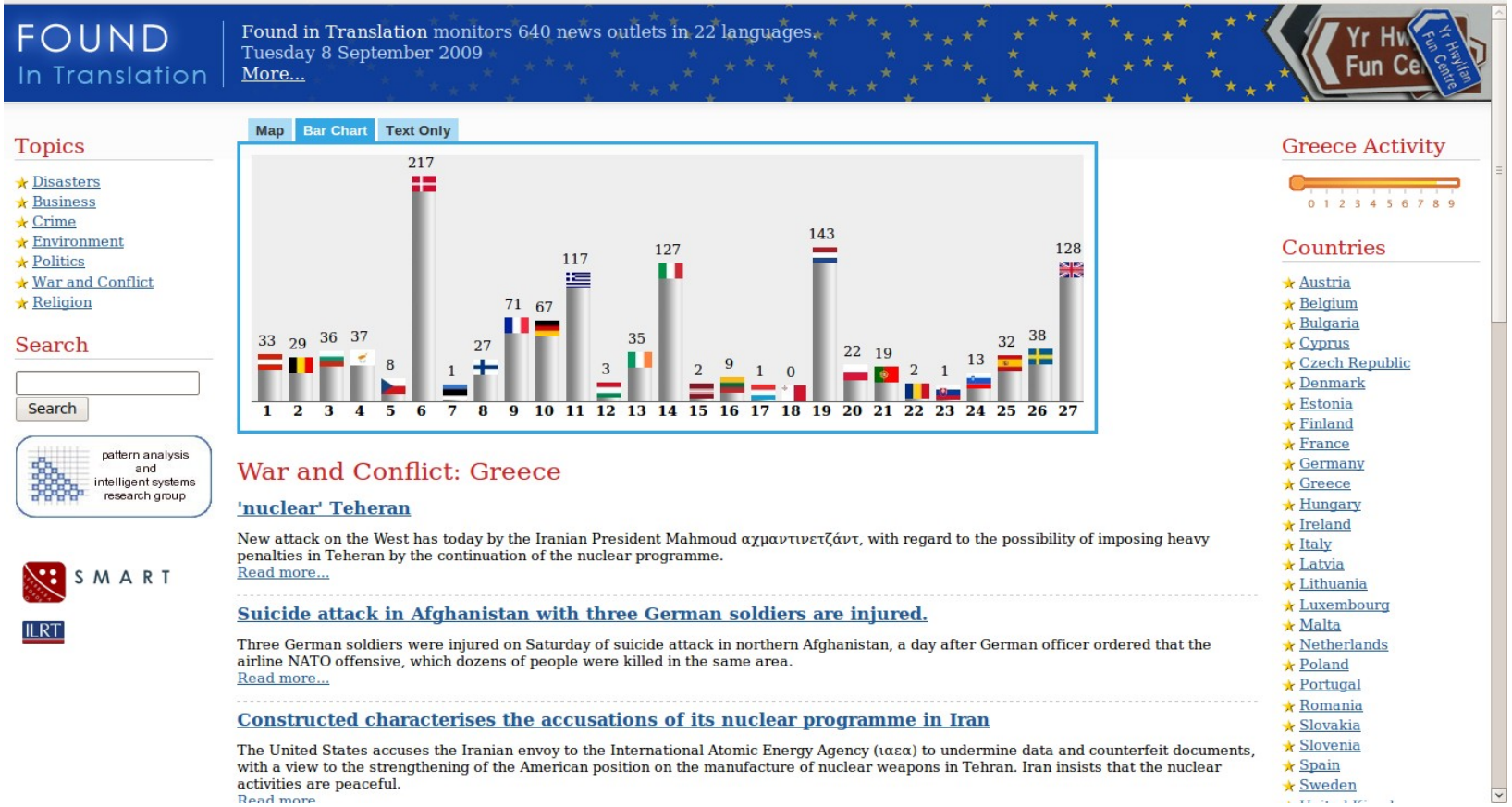
http://foundintranslation.enm.bris.ac.uk



University of BRISTOL

Found In Translation

http://foundintranslation.enm.bris.ac.uk



Found In Translation

For more information, please visit our poster.

Thank you!

University of BRISTOL

Found In Translation

Marco Turchi[1], Ilias Flaounas[2], Omar Ali[1], Tijl De Bie[1], Tristan Snowsill[1], Nello Cristianini[1,2]
[1]Department of Engineering Mathematics, Queen's Building, University of Bristol, Queen's Building, University Walk, Bristol, BS8 1TR,UK
[2]Department of Computer Science, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol, BS8 1UB, 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.

Web Interface
http://foundintranslation.enm.bris.ac.uk

Map View: Each country is coloured according to the levels of activity on a given topic. Green colours identify a weak interest for a particular topic, while red colours represent a strong interest.

Bar Chart View: The size of each bar represents the total number of translated and categorised news articles for each country for a particular topic.

1. Introduction

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 (FIT) is an online service that makes all this information available to English speakers under a common interface.

FIT **translates daily** news from 21 European languages into English.

A set of classifiers **categorises** each translated article into six topics, namely: "Accidents & Disasters", "Business", "Crime", "Environmental Issues", "Politics" and "War and Conflict".

All the news are **geo-located** and presented according to their topic on a "heat-map" of Europe to allow an easy comparison of the media attention on a given topic in the different countries of EU.

2. Statistical Machine Translation

We use **Moses** (http://www.statmt.org/moses/), a complete open source phrase based statistical machine translation toolkit.

For each language pair, an instance of Moses is trained using **Europarl** data and **JRC-Acquis Multilingual Parallel Corpus**.

The language model is refreshed daily using English news items of the last 15 days.

Translated articles of low quality are not displayed, but are still used to calculate the statistics presented in EU map and histograms.

3. Text Classification

Topic categorisation is performed using **Support Vector Machine (SVM)** classifiers. We use one classifier per topic we want to detect. We used the **SVM-Light** implementation of SVMs.

Each classifier is **trained off-line**, on a dataset of English articles which have been pre-labelled.

Linear kernels were used and the classifiers were trained with a choice of parameters aimed at achieving a **precision of above 90%:** Documents will be tagged only if we are very confident about the topic, while articles for which the confidence is lower will be left untagged.

4. Selecting representative news items

The statistics that are used to color the map are computed on a representative sample of news items at any given day. This is formed by selecting those articles that are in the main RSS feed of the top-10 news outlets per country. The ordering of the outlets is based on their **Alexa** rank score.

5. Conclusions

Found In Translation is an example of integration of state of the art technologies in machine intelligence, to enable a different user experience of the mediasphere: one where linguistic barriers can be overcome, and global trends in media content can be detected.

These trends are not possible to be monitored by a single human being.

University of BRISTOL