



Joint Research Centre

www.jrc.ec.europa.eu

Multilingual Named Entity Recognition and Name Variant Mapping

What is this about?

- We have developed multilingual text mining software for EMM (**Europe Media Monitor**; <http://emm.jrc.it/overview.html>), that automatically:
 - recognises named entities (persons, organisations, some events) in text;
 - detects variants referring to the same entity (e.g. *Javier Solana*, *Khavier Solana*, *خافيير سولانا*, *Хавьер Солана*, ...).
 - collects attributes about each person (profession, nationality, title, ...);
 - builds social networks (e.g. who gets mentioned with whom, who quotes whom).

Pervez Musharraf

Information about this person was last updated on Tuesday, September 15, 2009

Names	Key titles and phrases	External resources
Pervez Musharraf (eu,vi)	президент Пакистана (fr - 400)	
General Pervez Musharraf (dauv)	prezidento pakistanis (pt - 251)	
پرز مشرف (tr)	pakistanse president (nl - 215)	
Gen Pervez Musharraf (en)	pakistanse president (en - 181)	
Pervez Musarraf (tr)	prezidento pakistanis (es - 149)	
Pervez Musharraf (dauv)	president (de - 627)	
Pervez Musharraf (dauv)	president (da - 245)	
Gen Pervez Musharraf (en)	general (en,tr - 457)	
General Musharraf (dayen)	president (es,pt - 598)	
Pervez Musharraf (dauv)	president, gen (en - 82)	
Pervez Mywarraf (ru,uk)	prezident (de - 161)	
پرز مشرف (tr)	staatschef (de - 92)	
Pervez Musharraf (fr)	prezident (nl - 143)	
Pervezza Mularafa (pl)	prezident (fr - 312)	
Pervez Mywarraf (bg,ar)	prezidento pakistanis (it - 42)	
Pervez Musharraf (en,vi)	bagikan (tr - 231)	
Pervezza Musharraf (pl)	militärmachtgeber (de - 47)	



Image obtained automatically from Wikipedia

How does the software recognise name variants?

For all new names, compare to all known names, every day:

- Names in foreign language scripts (Arabic, Cyrillic, Greek) are first **transliterated**, using common transliteration schemes; e.g.
 - Greek: Κόφι Ανάν → Kofi Anan
 - Arabic: كوفي عنان → Kufi Anan
- Names are then **normalised** to a canonical form, exploiting empirically observed regularities;
- If the canonical form of the new name is the same as that of any of the known names and variants, two **string distance similarity metrics** are applied. The most similar are marked as variants.

Latin normalisation

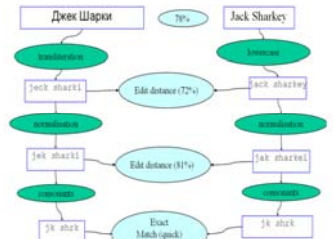
Normalisation rules:

- accented character → non-accented equivalent
- double consonant → single consonant
- ou → u
- 'al' →
- wl (beginning of name) → vl
- oe (end of name) → ov
- ck → k
- ph → f
- z → j
- š → sh
- x → ks

Remove vowels

Malik al-Saidoullaiev
 Malik al-Saidoullaiev
 Malik al-Saidoullaiev
 Malik Saidoullaiev
 ... mlk sdv

Name similarity calculation



How does the software recognise new names?

- If at least two uppercase words are found next to a **'trigger word'**;
- Trigger words are titles or professions (*president*; *teacher*), expressions relating to nationality, ethnicity or religion (*Thai*, *Berber*, *Catholic*), and more (*56-year old*, *has declared*, ...);
- Allows optional name parts inside the name, such as *van der*, *abu*, *bin*, *de la*, ... (e.g. *Osama bin Laden*);
- If a name part is a **known first name** (e.g. *Peter*, *Pierre*, *Pietro*, *Петер*, *بيتر*);

en	death of former Prime Minister Rafiq Hariri, blamed by many opposition
es	asesinato del expresidente ministro Rafic al-Hariri, que la oposición atribuyó
fr	l'assassinat de l'ex-dirigeant Rafic Hariri et le départ du chef de la diplom
nl	na de moord op oud-premier Rafiq al-Hariri gingen gisteren bijna een
de	libanesischen Regierungschef Rafik Hariri vor einem Monat wichtige B
sl	danjega libanonskega premiera Rafika Haririja. Libanonska opozicija si
et	möödumisele ekspeaminister Rafik al-Hariri surma põhjustanud pommpil
ar	اغتيال رئيس الوزراء السابق رفيق الحريري بالذمة يهودية وما حدث سابقاً
ru	Бывший премьер-министр Ливана Рафик Харирри, которого

Facts and Features

- Detects new names in **19 Languages** and recognises known names in all 50 EMM languages.
- Name database now contains almost **900,000 known names** plus about **275,000 variants**;
- Detects over **600 new names per day**, of which about 90 are recognised as variants of known names.
- Used in the EMM family of applications:



Languages

- ar - Arabic
- bg - Български
- es - Español
- de - Deutsch
- da - Dansk
- en - English
- et - Eesti keel
- fa - Farsi
- fr - Français
- it - Italiano
- nl - Nederlands
- no - Norsk
- pl - Polski
- pt - Português
- ro - Română
- ru - Russian
- sl - Slovenščina
- sv - Svenska
- tr - Türkçe

What is novel about this approach to name variant mapping?

- State-of-the-art:** The commonly used methods use long bilingual lists of names (e.g. lists of English names and their Arabic equivalent) to *learn* how letters and letter combinations are transliterated.
- Equivalences thus need to be learned for each language pair. For 19 languages, there are 171 language pairs!
- Our approach** represents all variants using one canonical form, making it possible to compare all languages with each other.

Why is this software useful?

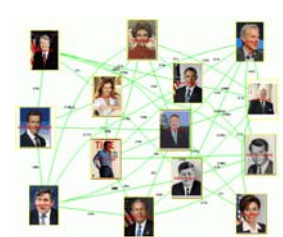
The name recognition and variant matching tools allow us to:

- display information on entities for each news article or cluster;
- collect more information about people from 19 languages;
- link related news across languages (we link news talking about the same names, locations and subjects);
- build social networks that are not biased by national viewpoints, but that are fed by multilingual multinational news.

Quotation Network



Co-occurrence network



Contributors

Ralf Steinberger, Bruno Pouliquen, Jenya Belyaeva, Hristo Tanev, Vanni Zavarella, Martin Atkinson, Brett Crawley, Erik van der Goot
 European Commission • Joint Research Centre
 Institute for the Protection and the Security of the Citizen
 Tel. +39 0332 785648 • Fax +39 0332 785154
 E-mail Format: Firstname.Lastname@jrc.ec.europa.eu

Selected publications

Pouliquen Bruno & Ralf Steinberger (2009). *Automatic Construction of Multilingual Name Dictionaries*. In: Cyril Goutte, Nicola Cancedda, Marc Dymetman & George Foster (eds.), Learning Machine Translation, pp. 59-76. MIT Press - Advances in Neural Information Processing Systems Series (NIPS).

Steinberger Ralf & Bruno Pouliquen (2009). *Cross-lingual Named Entity Recognition*. In: Satoshi Sekine & Ekaterina Bankhoff (eds.), Named Entities - Recognition, Classification and Use, Benjamin Current Topics, Volume 13, pp. 137-164. John Benjamins Publishing Company, ISBN 978-90-272-8922-3.

Pouliquen Bruno, Hristo Tanev & Martin Atkinson (2008). *Extracting and Learning Social Networks out of Multilingual News*. Proceedings of the social networks and application tools workshop (SocNet-08), pp. 13-16. Skalko, Slovakia, 19-21 September 2008.

Steinberger Ralf, Bruno Pouliquen & Erik van der Goot (2009). *An Introduction to the Europe Media Monitor Family of Applications*. In: Frédéric Gey, Noriko Kanda & Jussi Karlgren (eds.), Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-ELN 2009), pp. 1-8. Boston, USA: 23 July 2009.