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, خعیفر سولانا, Хавьер Солана, ...);
  - collects attributes about each person (profession, nationality, title, ...);
  - builds social networks (e.g. who gets mentioned with whom, who quotes whom).

How does the software recognise name variants?

For all new names, compare to all known names, every day:
1. Names in foreign language scripts (Arabic, Cyrillic, Greek) are first transliterated, using common transliteration schemes; e.g.
   - Greek: Κόφι Ακάν → Kofi Anan
   - Arabic: كوفي انان → Kufi Anan
2. Names are then normalised to a canonical form, exploiting empirically observed regularities;
3. 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

<table>
<thead>
<tr>
<th>Condition</th>
<th>Name similarity calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greek</td>
<td>0</td>
</tr>
<tr>
<td>Arabic</td>
<td>Kofi Anan (90%)</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>Кофи Акан (100%)</td>
</tr>
<tr>
<td>Greek</td>
<td>Κόφι Ακάν (100%)</td>
</tr>
</tbody>
</table>

Why is this software useful?

- 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.

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:

What is novel about this approach to name variant mapping?

- 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, Петер, پیتر)

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, Петер, پیتر)

Selected publications