Cross-lingual Similarity Calculation for Plagiarism Detection and More - Tools and Resources

Ralf Steinberger

European Commission – Joint Research Centre (JRC), Italy
Ralf.Steinberer@jrc.ec.europa.eu

Extended Abstract. A system that recognises cross-lingual plagiarism needs to establish – among other things – whether two pieces of text written in different languages are equivalent to each other. Potthast et al. (2010) give a thorough overview of this challenging task. While the *Joint Research Centre* (JRC) is not specifically concerned with plagiarism, it has been working for many years on developing other cross-lingual functionalities that may well be useful for the plagiarism detection task, i.e. (a) cross-lingual document similarity calculation, (b) subject domain profiling of documents in many different languages according to the same multilingual subject domain categorisation scheme, and (c) the recognition of name spelling variants for the same entity, both within the same language and across different languages and scripts. The speaker will explain the algorithms behind these software tools and he will present a number of freely available language resources that can be used to develop software with cross-lingual functionality.

The JRC's motivation to work on multilingual and cross-lingual language technology applications was to offer readers and news analysts access to up to 150,000 online media articles per day in about fifty languages that are gathered and processed by the fully automatic *Europe Media Monitor* (EMM) family of applications (Steinberger et al. 2009). The EMM applications are feely available online via the starting page http://emm.newsbrief.eu/overview.html. The automatic linking, across languages, of related news and of name variant spellings is performed in the EMM-NewsExplorer portal. NewsExplorer allows readers – starting from the news in one language – to jump directly to related news in the other languages, for 20 languages and all of its 190 language pairs. NewsExplorer furthermore collects information about named entities (mostly persons and organisations) from the many news articles in different languages and presents all the information on a single page, leveraging the knowledge about the entity variant spellings.

The name spelling variant recognition is performed daily for all newly identified entities in 20 languages, by first normalising the names to a simplified consonant signature and by then applying a string similarity measure between all new names and all previously known names in the JRC's name database. While other string distance similarity algorithms rely on learning equivalence rules from bilingual name lists (e.g. Knight & Graehl 1998), the JRC's hand-written normalisation steps and the string similarity calculation are the same for all languages. This makes it possible to efficiently merge the name variants found in many different languages and scripts without the need for training collections. For details on the algorithm, see Pouliquen & Steinberger (2009). The list of names and their spelling variants (linked via their unique numerical name identifier) have been released as *JRC-Names* (Steinberger et al. 2011).

The method to find related news clusters across the twenty NewsExplorer languages relies on simple cosine similarity calculations between four language-independent cluster representations: (a) a ranked list of EuroVoc subject domain codes; (b) a frequency list of the named entity mentions found in the clusters; (c) a frequency list of geo-references found and (d) a log-likelihood-weighted list of words found in the cluster. The first three representations are language-independent representations, i.e. they are numerical representations of normalised subject domains, persons, organisations and locations. Unlike the common cross-lingual text similarity measures, which are either based on translation and then using monolingual similarity measures or on bilingually trained vector space representations such as LSA (Landauer & Littman 1991) and KCCA (Vinokourov et al. 2002), the method used here is not limited to language pair-specific methods or resources. The tool that represents documents in 22 different languages as lists of EuroVoc subject domain codes has been released publicly under the name of JRC EuroVoc Indexer, JEX (Steinberger et al. 2012).

A number of further multilingual linguistic resources are available from http://langtech.jrc.ec.europa.eu/JRC_Resources.html.

Relevant references

Knight Kevin & Jonathan Graehl (1998). Machine Transliteration. Computational Linguistics 24:4, pp. 599-612.

Landauer Thomas & Michael Littman (1991). A Statistical Method for Language-Independent Representation of the Topical Content of Text Segments. Proceedings of the 11th International Conference 'Expert Systems and Their Applications', vol. 8: pp. 77-85

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, & Paolo Rosso (2010). Cross-language plagiarism detection. Language Resources and Evaluation 45 (1), 45-62.

Pouliquen Bruno, Ralf Steinberger & Camelia Ignat (2003). Automatic Identification of Document Translations in Large Multilingual Document Collections. In: Proceedings of the International Conference *Recent Advances in Natural Language Processing* (RANLP'2003). Borovets, Bulgaria, 10 - 12 September 2003.

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. 59-78. MIT Press - Advances in Neural Information Processing Systems Series (NIPS).

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

Steinberger Ralf, Bruno Pouliquen, Mijail Kabadjov & Erik van der Goot (2011). JRC-Names: A freely available, highly multilingual named entity resource. Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP'2011), pp. 104-110. Hissar, Bulgaria, 12-14 September 2011.

Steinberger Ralf, Mohamed Ebrahim & Marco Turchi (2012). JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool. Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012), pp. 798-805, Istanbul, 21-27 May 2012.

Vinokourov Alexei, John Shawe-Taylor, Nello Cristianini (2002). Inferring a semantic representation of text via cross-language correlation analysis . Advances of Neural Information Processing Systems 15.