Softcust

A Friendly Localized Platform for Multilingual Semantic Communication

Fabian Cretton, Zhan Liu and Anne Le Calvé

Institute of Business Information Systems
University of Applied Sciences of Western Switzerland
TechnoArk 3
3960 Sierre
SWITZERLAND
fabian.cretton@hevs.ch

RÉSUMÉ. Les dernières tendances de l'économie et de la mondialisation ont créés de nouveaux besoins pour améliorer la communication et la coopération dans les échanges internationaux. Cependant, il est souvent difficile de découvrir des informations pertinentes entre différentes langues et cultures dans les systèmes de gestion de l'information conventionnels. Dans ce papier, nous proposons un outil de communication sémantique et multilingue pour une gestion d'information automatisée dans un système convivial et localisé. Le modèle théorique est composé d'une ontologie multilingue basée sur Wikipédia pour fournir un moteur de recherche et un système de suggestion puissants. Notre solution actuelle comprend l'anglais, le français et le chinois.

ABSTRACT. The new economy and the globalization trends have brought new needs for increasing international business communication and cooperation. However, it is often difficult to discover relevant information across languages and cultures in conventional information management system. In this study, we have proposed a multilingual semantic communication for automated information management through a user-friendly localized system. The theoretical model is made of a multilingual ontology based on linguistic knowledge from Wikipedia in order to provide a powerful search engine and suggestion system. Our current solution includes English, French and Chinese.

MOTS-CLÉS : communication sémantique multilingue, gestion des données et des connaissances, ontologie multilingue, moteur de recherche sémantique, plate-forme de business international.

KEYWORDS: multilingual semantic communication, knowledge and data management, multilingual ontology, semantic search engine, international business platform.

Actes du 5e Colloque National sur la Recherche en Informatique et ses Applications – Ziguinchor, Sénégal, Avril 2013

1. Introduction

Given the strong demand of the Chinese markets for professional management software, we have conducted applied research on localizing Swiss-made computer programs on the Chinese market during the Sino-Swiss Software Customization Network Project[1]. This research included several technical topics such as intellectual property protection, Chinese characters display or ePayment. This previous project was successful in localizing a few products for the Chinese market, but some repetitive research tasks could be improved and partly automated, especially for matching the Swiss-made products with the existing demands on the Chinese market. Therefore we present this SoftCust project with the main goal to develop an intelligent matching platform to help the Chinese and Swiss IT SMEs match their competences, needs, offers and joint activities intelligently.

As stated by the W3C, the Semantic Web «provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries»[2]. Nowadays semantic web knowledge's representation (ontologies) is a popularized mean of designing and sharing a domain model and its data. This is true about any domain (e.g. health care, finance or business), and also about representing a language's lexical semantic as well as cross-language information. Moreover, semantic tools provide powerful ways for men and machines to process data, for instance using specific features of the query language or using rules engines (i.e. reasoners) to infer new knowledge and perform data validation checks.

Therefore our research question is: how can semantic web technologies improve the performance of an exchange platform in designing a friendly localized system with multilingual semantic communication.

In the rest of the paper we address our research question using a design science approach and we structure our paper according to [12] guidelines: the first part discusses some related work for the semantic matching aspect. In the second part we present the chosen solution and then its implementation in the third part. Some discussion and evaluation is presented in the fourth part. We then conclude with some future directions, improvements and challenges to overcome in the fifth part.

2. Related Work

To achieve semantic matching in a multilingual exchange platform, we mainly

[1] http://iet.hevs.ch/valais/softcust-sino-suisse-software-customization.html
[2] http://www.w3.org/RDF/FAQ

tackled two topics: text mining and cross-language information retrieval.

Text mining is a domain of natural language processing (NLP), a computer science field where machines extract knowledge from text written in natural language. In most of NLP tasks machine learning tools have been successfully used for years. Downsides of that approach are the amount of human work to annotate existing corpus in order to train the system and the limitation of the system that becomes efficient only for the domain it has been trained on. Recently, knowledge based technics are giving comparatively good results, and even better results in some cases, as demonstrated by [11] for word-sense disambiguation or [4] for text classification. In this approach, the knowledge is described in a structured manner, for instance using ontologies, and the NLP algorithms exploit that knowledge to perform their tasks on the texts.

Ontologies are becoming a common way to describe knowledge. They are based on W3C standards as RDF and OWL, or the SKOS[3] vocabulary to define concept schemes, glossaries and terminologies. NLP tools making use of ontologies are spreading, especially for named entity recognition. Those tools can thus recognize entities (e.g. persons, organizations or locations) with very good results as presented in [8][13], sometimes as online services[4]. However, the core of our problematic relies more on dealing with common words (synonyms for instance) then named entities.

Information retrieval (IR) technics have been elaborated to identify the most relevant elements of a corpus to satisfy a search. Full-text search, providing a simple match between a search key-word and that word in the indexed texts, is quite limited when it comes to lexical semantics. It can be enhanced using query expansion[5] techniques which, as stated by [2], are becoming quite mature but still face difficulties. The same conclusions can be drawn for cross-lingual methods (CLIR), a field which produces good results but still considered a non-trivial task [14][5]. Traditional NLP and IR technics both have to deal with lexical semantic and cross-lingual issues, for which semantic web technologies provide very interesting features. Our view of the semantic information retrieval issue is similar to the proposals of KIM [13] [17] and TAP [18], which focus on automatic population and annotation of documents. However, our solution includes Chinese and is more specifically tailored to the case.

We need to handle full-text search enhanced with semantics and cross-lingual capabilities for 3 specific languages: French, Chinese and English. This requires an ontology that is generic enough to contain the three language's lexical semantic such as synonymy, hyponymy or holonymy, about terms specific to the computer science domain, but that could be easily extended to any domain. A number of generic

[3] http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/
[4] http://viewer.opencalais.com/, http://www.alchemyapi.com/
[5] http://en.wikipedia.org/wiki/Query_expansion

ontologies are available as OpenCyc[6], UMBEL[7], or Wikipedia based as DBPedia [1] or Yago [15][9]. Unfortunately, even though those resources do contain multilingual information, the languages we need are not well described and also lack lexical semantic. Other ontologies as WordNet do contain rich linguistic information, but there is currently no multilingual version of WordNet that includes French, English and Chinese in a freely available resource.

As we can see, many solutions propose the features related to our platform, but there is no integrated solution that can carry out all the tasks for French, English and Chinese. We thus needed to develop new tools and ontologies to answer our research question.

3. Solution

We designed a platform that helps users with different roles (programmer, distributor, vendor, etc.) to find a match (a user looking for a software, a vendor looking for a distributor, etc.) based on some structured data and also descriptions in natural language. The platform is a hybrid system based on a relational database and a triple store to manage classical and semantic knowledge. The semantic search is also a compound engine based on semantic indexing of text with an ontology, coupled with a classical full-text search.

A generic multilingual ontology has been created from the English, French and Chinese Wikipedia. NLP functionalities are used to analyze users' descriptions and link them with the ontology terms. Then, cross-lingual matches are performed between the requests and the offers. A main component of our platform, called «semText», allows semantic indexing of texts and then performs matches based on cross-lingual and lexical semantic defined in the ontology. We did generate a SKOS ontology from Wikipedia dumps, using the Java-based Wikipedia Library (JWPL) as presented in [16]. That solution has the advantage to be easily extended to over 200 languages found in Wikipedia, and it can also be enriched by other related information found in Wiktionary[8] or in the LOD[9] for instance. The resulting multilingual ontology does contain the needed information about a word: its translation in the 3 languages, synonyms, and exploitable links to other concepts (e.g. hyponyms, related words).

To handle text, NLP systems have to carry out pre-processing tasks [10] as tokenization, stemming/lemmatization, part-of-speech tagging, for which a number of tools are available. We did choose the renowned GATE architecture [3] available as a

[6] http://sw.opencyc.org/
[7] http://www.umbel.org/
[8] http://www.wiktionary.org/
[9] http://linkeddata.org/

user interface and a java API. Having the ontology on one side and the text descriptions on the other, semText uses GATE to generate a semantic lookup graph of the texts, i.e. link the words of a text to the concepts of the ontology. To describe the lookup graph we designed the OntoManagNLP ontology. Here is some information about the different components we used for GATE. The LKB (Large Knowledge Base) Gazetteer from OntoText handles the very big French and English gazetteers, representing millions of concepts. The classical GATE's ANNIE gazetteer was used for about 200'000 Chinese concepts, the LKB being not able to handle Chinese words. Tokenization being particularly subtle for the Chinese characters, we added the ICTClas[10] library to both, Lucene and Gate, as the default tools were giving poor results. Finally, Tree-tagger[11] is used for the French part-of-speech.

The texts are thus indexed with the ontology, but also with Apache Lucene[12] for full-text search. Using the ontology's links between words to search for the texts, allows for more precise querying than using a traditional query expansion technique. When a user looks for a word, the ontology is used to look for texts which have lookups to any related word (e.g. synonym, hyponym), in a very precise manner, moreover handling directly the cross-lingual search. The need for full-text search was that no matter how generic the ontology is, it will never contain all words of a language. However, words in the indexed texts but not in the ontology must still be usable for searches even though there is no lookup for them. We thus created a hybrid search: semantic search for the keywords (or related words) found in the ontology, and full-text search for the others (handling translation on the fly for multi-lingual results).

Based on the relevant literatures, we create an artifact in the form of a model [7] to represent our multilingual semantic search engine. We adopt a design science research methodology and refer to existing guidelines for design theories [6]. In order to better understand our solution, we present a simple scenario. Suppose that Alice is a manager in a small CRM software company in Switzerland. Because of the financial crises and the increasing competition in domestic market, she is looking overseas for potential clients. Ming is running his textile company in China and is willing to buy CRM software to help him understanding the needs of its customers. However, he notices that CRM software is less mature in China and wants localized CRM software from outside of China. Unfortunately, there are some problems to get Alice and Ming connected, and that's where our solution comes into play. When Ming uses this platform with keyword search «client management software» in Chinese, he can have a list of «CRM» software companies and their product description. Therefore, our solution provides cross-lingual matching enhanced with related words handling (synonyms, generalization and specialization) in the context of a specific search.

[10] http://code.google.com/p/openictclas/
[11] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[12] http://lucene.apache.org/core/

4. Implémentation

In the Intelligent Matching Platform users create their, describe their domain of activities and upload documents to add information on their products. The Softcust system manages classic multilingual information with a MySql database. Information is additionally indexed semantically and available for search.

Partners' searches are based on several criteria: keywords, roles, domain types, countries and languages. The system performs the semantic matching and displays the results in the user's mother tongue (translation being done with the Google Translate API[13]). Searches can be saved as wish lists, and thus managed by a notification system that will notify the users whenever new corresponding entries enter the system.

The back-end triple store is OWLim 4.3 from OntoText[14], chosen because it supports loading/querying/reasoning on a large amount of triples, as well as very useful SPARQL 1.1 features.

5. Discussions and Conclusions

An evaluation has been done by the SoftCust team with real data coming from their former project. They did appreciate that the underlying semantic tool works transparently on the existing texts, without having to enter new data. Results are thus very encouraging as the tool does allow performing automatic match between data in different languages and using different vocabulary or expressions.

In this research project we realized a semantic enriched platform to help Swiss and Chinese actors of the computer science field to find matches between supply and demands. For that system based on semantic web technologies we've generated a SKOS ontology from Wikipedia that includes lexical semantic of the French, English and Chinese languages. Our research question was thus answered as the semantic engine allows richer search results than simple full-text search, handling synonyms and related words (as hyponyms) in a more efficient way then traditional IR query expansion, and last but not least, with cross-lingual capabilities. For the scientific community, we provide new alternatives for international communications and business models into the management of international "buyer-seller" relationship.

In a future work, we would like to improve the evaluation of the results by implementing some IR techniques, and enrich the ontology in the different languages, adding information from the new versions of Wikipedia and complement it with

[13] https://developers.google.com/translate/
[14] http://www.ontotext.com/

Wiktionary or any data source from the LOD. SemText will be used to work on the lookup graph and implement technics of word sense disambiguation as well as text classification. Moreover, the main remark about further tests and downsides evaluation is that the system needs to contain a large amount of descriptions and searches to see how it performs in a real-time environment.

We thank Ontotext who supported us with a research license for OWLIM.

6. Bibliography and biography

6.1 Bibliography

[1] Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R. and Ives Z., *DBpedia: a nucleus for a web of open data*. In Aberer et al. (Eds.): The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea. Lecture Notes in Computer Science 4825 Springer 2007, ISBN 978–3-540–76297–3, 2007

[2] Carpineto C. and Romano G., *A survey of automatic query expansion in information retrieval*. ACM Computing Surveys, 44(1), 2012

[3] Cunningham H., *GATE, a General Architecture for Text Engineering*. Computers and the Humanities, 2002

[4] De Maat E., Krabben K., and Winkels R., *Machine Learning versus Knowledge Based Classification of Legal Texts*. In Proceedings of the 2010 conference on Legal Knowledge and Information Systems: JURIX 2010: The Twenty-Third Annual Conference, Radboud G. F. Winkels (Ed.). IOS Press, Amsterdam, The Netherlands, 87-96, 2010

[5] Dolamic L. and Savoy J., *Monolingual and Bilingual Searches: Evaluation, Challenges and Failure Analysis*. Submitted, 2008

[6] Gregor S. and Jones D., *The anatomy of a design theory*. Journal of the Association for Information Systems, 8(5), pp. 312–335, 2007

[7] March S. T. and Smith G. F., *Design and natural science research on information technology*. Decision Support Systems, 15(4), pp. 251-266, 1995

[8] Mendes P. N., Jakob M., García-Silva A., and Bizer C., *DBpedia spotlight: shedding light on the web of documents*. In Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11), ACM, New York, NY, USA, 1-8, 2011

[9] Müller C. and Gurevych I., *Using Wikipedia and Wiktionary in domain-specific formation retrieval*. In Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access (CLEF'08), 219-226, 2008

[10] Nadkarni P., Ohno-Machado L., Chapman W. W., *Natural language processing: an introduction*. Journal of the American Medical Informatics Association, 18: 544-551, 2011

[11] Navigli R., *Word Sense Disambiguation: A Survey*. ACM Computing Surveys, 41(2), ACM Press, pp. 1-69, 2009

[12] Peffers K., Tuunanen T., Rothenberger M. A., and Chatterjee S., *A Design Science Research Methodology for Information Systems Research*. Journal of Management Information Systems, 24(3), 45-77, 2007

[13] Popov B., Kirayakov A., Ognyanoff D., Manov D., Kirilov A, *KIM - a semantic platform for information extraction and retrieval*. Natural Language Engineering 10 (3/4), pp. 375-392, 2004

[14] Savoy, J., *Comparative study of monolingual and multilingual search models for use with Asian languages*. ACM TALIP, vol. 4(2), pp. 163–189, 2005

[15] Suchanek F. M., Kasneci G., Weikum G., *Yago: A large ontology from Wikipedia and wordnet*. Journal of Web Semantics 6 (3) 203-217, 2008

[16] Zesch T., Müller C. and Gurevych I., *Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary*. In Proceedings of the Conference on Language Resources and Evaluation LREC, electronic proceedings, 2008

[17] Kiryakov A., Popov B., Terziev I., Manov D., and Ognyanoff D., *Semantic Annotation, Indexing, and Retrieval*. J. Web Semantics, vol. 2, no. 1, pp. 49-79, 2004.

[18] Guha R. and McCool R., *TAP:a Semantic Web platform*. Computer Networks, 42(5), pp. 557–577, 2003.

6.2 Biography

The authors work at the University of Applied Sciences of Western Switzerland.

Fabian Cretton is an active analyst and senior developer specialized in semantic web technologies in various domains such as user modeling, adaptive interfaces, ontology design, knowledge management and natural language processing.

Zhan Liu is also PhD candidate at the Department of Information Systems of the University of Lausanne, Switzerland. His main research interests are semantic web, privacy and security management, mobile business and mobile technology development. His work has been published in journals, book chapters and conferences.

Dr. Anne Le Calvé is professor since 1999. She holds a PhD in computer science - about meta search engine methodologies in information retrieval domain. Her research activities focus on knowledge management, information modeling and semantic web technologies, applied to different fields such as eTourism, eGov, personal information management and user modeling.