Expert Systems With Applications 57 (2016) 248–269

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A semantic framework for textual data enrichment

Yoan Gutiérrez, Sonia Vázquez*, Andrés Montoyo

Department of Software and Computing Systems, University of Alicante, Spain

ARTICLE INFO

Article history:
Received 23 December 2015
Revised 29 February 2016
Accepted 25 March 2016
Available online 30 March 2016

Keywords:
Recommender systems
Framework
Integrated semantic resources
Sentiment analysis
Word Sense Disambiguation
Content categorisation

ABSTRACT

In this work we present a semantic framework suitable of being used as support tool for recommender systems. Our purpose is to use the semantic information provided by a set of integrated resources to enrich texts by conducting different NLP tasks: WSD, domain classification, semantic similarities and sentiment analysis. After obtaining the textual semantic enrichment we would be able to recommend similar content or even to rate texts according to different dimensions. First of all, we describe the main characteristics of the semantic integrated resources with an exhaustive evaluation. Next, we demonstrate the usefulness of our resource in different NLP tasks and campaigns. Moreover, we present a combination of different NLP approaches that provide enough knowledge for being used as support tool for recommender systems. Finally, we illustrate a case of study with information related to movies and TV series to demonstrate that our framework works properly.

1. Introduction

Recent advances in modern technologies have motivated the development of different techniques to improve human–machine communication. Internet and new communication tendencies such as: short messages, forum participations, social networks, etc., have led to a revolution in the way in which people work, communicate and manage their free time. As a consequence of this technological revolution, a huge quantity of information is generated in different social contexts via diverse sources such as: forums, blogs, microblogs, social networks, etc. As a result, people are able to share their knowledge, expectations and emotions through Internet and they may also influence political, economic or social behaviour. At this point, governments, enterprises or even celebrities need to manage this information in order to extract relevant knowledge, social tendencies, etc. Because of this new context, research community in Natural Language Processing (NLP) have developed different tools with which to analyse news and opinions in order to discover what people think or how they perceive past, present and future.

At present, personalisation and recommender systems have gained popularity. In fact, recommender systems began to appear in the market in 1996 (Udi, Ash, & John, 2000). Since then, several approaches have been developed (Gediminas & Alexander, 2005):

- Content-based: these systems try to find products, services or contents that are similar to those already evaluated by the user. In this kind of system, user's feedback (that can be collected in many ways) is essential to support and accomplish recommendations (Marco De, Pasquale, Giovanni, & Pierpaolo, 2008).
- Knowledge-based: these systems model the user profile in order to, through inference algorithms, identify the correlation between their preferences and existing products, services or content (Walter, Maria Luisa, Rafael, & Francisco, 2012).
- Collaborative filtering: these systems create/classify groups of users that share similar profiles/behaviours in order to recommend products, services or content that has been well evaluated by the group to which a user belongs (Perner, Candillier, Meyer, & Boulle, 2007).
- Hybrid: these systems combine two or more techniques previously mentioned to improve the "quality" of recommendations (Shinde & Kulkarni, 2012).

Dealing with textual information and obtaining valuable knowledge require advanced natural language techniques to solve different kinds of problems: document correction, automatic translation, summary elaboration, opinion extraction, Word Sense Disambiguation, etc. Solving all of these problems requires a considerable linguistic knowledge and, even more importantly, a high computational cost.

In the vast majority of tasks in NLP it is necessary to use external resources such as: Machine-readable dictionaries,[1] Thesaurus,[2] Ontologies[3] and others. These resources have different in-

* Corresponding author. Tel.: +34 965903772; fax: +34 965909326.
E-mail addresses: ygutierrez@dlsi.ua.es (Y. Gutiérrez), svazquez@dlsi.ua.es, svazquez@dlsi.ua.es (S. Vázquez), montoyo@dlsi.ua.es (A. Montoyo).

http://dx.doi.org/10.1016/j.eswa.2016.03.048
0957-4174/

[1] Dictionaries of words available in electronic format.
[2] Provides relationships among words (i.e., synonyms, antonyms and others).
[3] Conceptualisation of a domain in order to share information among different agents.

ternal structures, interfaces, concept relations and other characteristics. One of the most frequently used resources in its different versions is WordNet[4] (WN) (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). Various semantic resources related to WordNet have consequently been developed in different domains or by using semantic integration. But it is still difficult to find resources that provide semantic integration in different domains and which are useful for specific NLP tasks.

In this work we present a new semantic resource (ISR-WN) and a set of different methods to take advantage of it with the aim of enriching texts with semantic information. As a result, we provide a semantic framework suitable of being used as support tool for content-based recommender systems by annotating texts with different features such as: sentiments, polarities or domain labels. In order to analyse the results of the semantic enrichment process, we have carried out a comprehensive case of study using texts from movies and TV series reviews obtained from IMDb. [5] Finally, we have evaluated how our proposed framework works comparing our results with real ratings.

To summarise, we point out the main contributions of this work:

- Taking advantage of a semantic resource with different dimensions previously developed (ISR-WN).
- The use of a set of NLP methods based on ISR-WN to take advantage of each one of its semantic dimensions.
- Providing a new semantic framework that is able to enrich texts in several dimensions with the aim of obtaining a support tool for content-based recommender systems.
- An exhaustive evaluation with real datasets to demonstrate how it works.

The document is structured as follows. After this introduction, each semantic resource used in ISR-WN is described, and an in-depth analysis of the different approaches for semantic integration resources in NLP is also presented. Having evaluated previous proposals, in Section 3 we go on to show how ISR-WN was developed. An evaluation according to its integration effectiveness is then provided in Section 4. In Section 5 we provide a brief description of the different NLP tasks selected to enrich texts. Section 6 describes the characteristics of a case of study to illustrate how our framework works with real data obtained from IMDb. In Section 7 we show some examples of how the semantic enrichment approaches are used to annotate texts. Section 8 provides the experiment results of the case of study and Section 9 presents a discussion about the results obtained. Finally, the conclusions and further works are presented in Section 10.

2. Related work

This section presents the different semantic resources that are integrated into ISR-WN and a comparison with other semantic integration resources.

2.1. WordNet

As mentioned in the previous section, WN is one of the most frequently used semantic resources in computational linguistics (Navigli, 2009). WN is a lexical database for the English language also considered as ontology. It was created at the University of Princeton[6] and it represents a semantic conceptual and structured network of nouns, verbs, adjectives and adverbs. The basic unit of knowledge is the synset (synonym sets), which represents a lexical concept (Ševčenko, 2003). A synset is associated with a unique eight-digit number called an offset (this number is the position in the data file). Each synset represents different senses which are related through the use of semantic, conceptual or lexical connections. The result of this set of connections is a wide navigable network with a high number of interrelations among different word senses.

The semantic relations among synsets are:

- *Synonymy*
- *Antonymy*
- *Hyponymy/Hyperonymy*
- *Meronymy/Holonymy*
- *Entailment and cause,*
- *and others…(more details in*[7]*)*

WN establishes the frequency of usage of each word sense (synset)in its internal relationships. [8] For example, the word *image* has eight senses in WN 2.0 (see Table 1). As we can observe, one word has different senses, there is a sentence (gloss) which describes each one and each sense has a set of synonyms that are ordered by their frequency of usage.

It is important to emphasise that WN has been adapted to different languages: English, Spanish, Dutch, Italian, German, French, Czech, Estonian, Swedish, Norwegian, Danish, Greek, Portuguese, Basque, Catalan, Romanian, Lithuanian, Russian, Bulgarian, Slovenian and others that are under development. These versions have been developed under the supervision of the University of Princeton and later under that of the Global WordNet Association.[9]

This research work is based on two versions of WN: WN 1.6 with 99,643 synsets, of which 66,025 are nouns, 17,915 are adjectives, 3575 are adverbs and 12,127 are verbs, and WN 2.0 with 115,424 synsets, of which 79,689 are nouns, 18,563 are adjectives, 3664 are adverbs and 13,508 are verbs.

2.2. Semantic resources aligned to WordNet

Owing to the fact that WN has been used in many NLP research works, a set of different semantic resources aligned to WN synsets has been developed with the aim of obtaining more knowledge. Some of these resources were created from WN, such as: WordNet Domains[10] (Magnini & Cavaglia, 2000), WordNet Affect[11] (Magnini & Cavaglia, 2000; Sara & Daniele, 2009) and Semantic Classes[12] (Izquierdo, Suárez, & Rigau, 2007). Others emerged from the association of pre-produced tags, i.e., SUMO.[13]

The resources used in our proposed semantic integration resource (ISR-WN) are described in detail below.

2.2.1. WordNet Domains

This is a resource for the English language. WordNet Domains (WND) includes a set of Subject Field Codes (SFC) (Magnini & Cavaglia, 2000) with which to enrich WN synsets. Each SFC groups a set of words related to the same domain. On the one hand, these domains identify the context of the definition and on the other, they allow a quick search of concepts to take place. For example, if we are searching for the meaning of *disc* in the Computer Science context, we need only check the domain label preceding each definition (in this case, *Computer Science*) until we find the correct

[4] http://wordnet.princeton.edu/.
[5] http://www.imdb.com.
[6] http://wordnet.princeton.edu/wordnet/.
[7] http://wordnet.princeton.edu/man/wninput.5WN.html.
[8] https://wordnet.princeton.edu/man/cntlist.5WN.html.
[9] http://www.globalwordnet.org/.
[10] http://wndomains.fbk.eu/.
[11] http://wndomains.fbk.eu/wnaffect.html.
[12] http://rua.ua.es/dspace/bitstream/10045/2522/1/ranlp07BLC2.pdf.
[13] http://www.ontologyportal.org/.