Concept-based Item Representations for a Cross-lingual Content-based Recommendation Process

Fedelucio Narducci, Pierpaolo Basile, Cataldo Musto, Pasquale Lops
Annalina Caputo, Marco de Gemmis, Leo Iaquinta, Giovanni Semeraro

[a] *Department of Computer Science, University of Bari Aldo Moro*
Via E. Orabona 4, I-70125 Bari, Italy

Abstract

The growth of the Web is the most influential factor that contributes to the increasing importance of text retrieval and filtering systems. On one hand, the Web is becoming more and more multilingual, and on the other hand users themselves are becoming increasingly polyglot. In this context, platforms for intelligent information access as search engines or recommender systems need to evolve to deal with this increasing amount of multilingual information. This paper proposes a content-based recommender system able to generate cross-lingual recommendations. The idea is to exploit user preferences learned in a given language, to suggest item in another language. The main intuition behind the work is that, differently from keywords which are inherently language dependent, concepts are stable across different languages, allowing to deal with multilingual and cross-lingual scenarios. We propose four knowledge-based strategies to build concept-based representation of items, by relying on the knowledge contained in two knowledge sources, i.e. Wikipedia and BabelNet. We learn user profiles by leveraging the different concept-based representations, in order to define a cross-lingual recommendation process. The empirical evaluation carried out on two state of the art datasets, `DBbook` and `Movielens`, shows that concept-based approaches are suitable to provide cross-lingual recommendations, even though there is not a clear advantage of using one of the different proposed representations. However, it emerges that most of the times the approaches based on BabelNet outperform those based on Wikipedia, which clearly shows the advantage of using a native multilingual knowledge source.

Keywords:
Content-based Recommender Systems, Concept-based representations, Wikipedia, BabelNet

1. Introduction

In 1998, 70% of the content on the Web was in English [35]. Nowadays about 45% of the websites provides content in a language different from English and the number

Email address: `name.surname@uniba.it` (Fedelucio Narducci, Pierpaolo Basile, Cataldo Musto, Pasquale Lops
Annalina Caputo, Marco de Gemmis, Leo Iaquinta, Giovanni Semeraro)

DOI: 10.1016/j.ins.2016.09.022

of non-English pages is rapidly growing[1]. In the past, multilingual websites were in a small number due to the high costs of development and maintenance. Companies could hardly afford those costs also because the number of non-English Internet users was really small and the potential revenues did not justify the required investments [49]. However, the rapid growing of non-English Internet users is changing that scenario. In a recent statistics updated on June 30, 2015, users with the largest growth of the Internet use in the period from 2000 to 2015 are Arabic speakers (+6,091%), Russian speakers (+3,227%), Chinese speakers (+2,080%), whereas English speakers (+505%), Germans speakers (+204%), and Japanese speakers (+144%) occupy the last positions[2]. Accordingly, we can state that the Web is becoming more and more multilingual, with the top websites, such as Bing, Google, Wikipedia, etc., offering their content in hundreds of languages.

Another relevant aspect is that users themselves are becoming increasingly polyglot, i.e. people are increasingly proficient in more than one language [48]. It has been estimated that more than half of the world population is bilingual [22], while statistics about language education in the European Union (in 2012) show that on average 94.5% of secondary education pupils now learn English in general programs, and 50.6% learn two or more languages[3].

According to this scenario, platforms for intelligent information access as *search engines* or *recommender systems* need to evolve in order to effectively deal with this increasing amount of multilingual information. Indeed, information retrieval (IR) systems may allow to retrieve relevant results in a language different from that used to issue the query, while information filtering (IF) systems may suggest interesting items in a language different from that the user explicitly used to express her interests. This problem is known in the literature as Cross-lingual Information Access.

This clearly motivates the need for efficient and effective IF and IR techniques that cross the boundaries of languages. In that context, we must face with the so-called *vocabulary mismatch* problem [50], i.e. relevant documents might potentially be judged as irrelevant due to a low textual overlap between query and document, or interesting items might be judged not interesting due to the low overlap between the user profile and the item descriptions. An extreme case of the vocabulary mismatch problem arises in settings where relevant (interesting) documents are written in other languages than the one of the query (user profile) [47]. One way to overcome the language barrier is to focus on the *concepts associated to words*, i.e. their *meaning*. The meaning of words is inherently multilingual, since concepts remain the same across different languages, while words used to describe those concepts in each specific language change. A concept-based representation of items and user profiles could represent an effective way to have a language-independent representation, which could act as a *bridge* among different languages.

In this paper, we investigate whether a concept-based representation is an effective strategy to provide language-independent representations of items and user profiles, which in turn allows an effective cross-lingual content-based recommendation process.

In this paper we aim at answering to the following research questions:

[1] w3techs.com/technologies/overview/content_language/all

[2] www.internetworldstats.com/stats7.htm

[3] ec.europa.eu/eurostat/statistics-explained/index.php/Foreign_language_learning_statistics

- **R1**: Are knowledge-based strategies able to face the content-based cross-lingual recommendation problem?
- **R2**: Are concept-based representations able to provide effective content-based cross-lingual recommendations compared to translation-based approaches?
- **R3**: Are knowledge-based representations effective to provide content-based cross-lingual recommendations when limited textual content is available?

To answer to these questions, we have performed an in-depth experimental evaluation on two state of the art datasets, i.e. `DBbook` and `MovieLens`, in order to assess the effectiveness of the cross-lingual recommendations by taking into account concept-based representations obtained by leveraging two different knowledge sources, i.e. Wikipedia and BabelNet [38], different languages, and item descriptions of different length. The results show that concept-based approaches which abstract from surface representations are suitable for cross-lingual scenarios. A clear advantage of using one of the proposed approaches did not emerge, although the use of a native multilingual knowledge source such as BabelNet often leads to better results with respect to the use of Wikipedia. Furthermore, processing shorter item descriptions leads to better results as well.

The article is organized as follows. Section 2 discusses the related work, while Section 3 describes the cross-lingual content based recommendation process. The adopted knowledge-based strategies to build language independent concept-based representations are described in Section 4. Finally, experimental results are shown in Section 5, and the conclusions are drawn in Section 6.

2. Related Work

Multilingual Information Access (MLIA) and Cross-Lingual Information Access (CLIA) are the most relevant tasks for the research presented in this work. MLIA is defined as the problem of accessing, querying and retrieving information from collections in any language and at any level of specificity [42]. MLIA incorporates CLIA, which refers to technologies used for accessing a data collection in a target language l_2, by using a source language l_1, where $l_1 \neq l_2$. MLIA and CLIA have been widely investigated in the literature, in particular in the Ontology Matching and IR research areas. To the best of our knowledge, the topic of Cross-Lingual and Multilingual Information Filtering has not been properly investigated in the literature yet. However, it is known that Information Filtering and Information Retrieval have common roots [5] and for this reason in this Section we start to analyze researches related to Cross-Language Information Retrieval (CLIR).

The literature on CLIR is very rich. Two main categories of approaches can be identified: *translation-based* approaches and *concept-based* ones.

2.1. Translation-based approaches

Most research activities address CLIR by performing a preliminary translation process, which can concern the document collection [40] or more simply the query [11]. In the first case, the whole collection is translated in all the languages the query can be

formulated in. The main advantage of this approach is that there is no speed penalty caused by translation at query time. However, this solution is really time consuming and might be necessary to periodically redone due to the evolution of the translation algorithms over the years. Furthermore, any document-translation approach requires to define in advance the languages in which the query can be formulated and, in addition, to store all the translated versions of the collection. The second approach consists in translating the query in the languages used for representing the documents. The query is translated *on-the-fly* and that entails performance penalty. The major problem is the sense disambiguation: the query is usually short, hence the correct translation of each term is a complex task. However, the user could be capable of understanding the translation of the query, and correct it before the use [39]. This approach demonstrates a higher flexibility with respect to the document-translation approach.

Several researchers compared the two approaches by using the same translation tool. In [17], query and document translation approaches are compared using the IBM translation based system, but a clear advantage of an approach with respect to another did not emerge. McCarley [31] demonstrated that the effectiveness is more influenced by the translation direction (e.g. Italian-to-English, English-to-Italian) rather than the decision of translating queries or documents. That result demonstrates that a crucial role is played by the translation process.

In the meanwhile, machine translation algorithms have drastically improved their performance. Statistical Machine Translation algorithms proved to be more effective with respect to other approaches (e.g. rule-based) [41] and are now widely adopted by the major machine translation tools (e.g. Google Translate[4], Microsoft Translator[5]). Accordingly, bilingual experiments of CLEF 2009[6] obtained an effectiveness up to 99% of monolingual baseline, mostly using the Google Translate service [42].

In this article we evaluated translation-based approaches for providing cross-lingual recommendations. More specifically, we employed Bing[7] to bridge the gap between different languages. The translation-based approach has been combined with concept-based approaches described in the next section.

2.2. Concept-based approaches

A third translation-based approach exploits a pivot language [46, 21] to obtain a common document representation. In that case, a direct translation from the source language to the target one is not performed, but the two languages are represented in a third common language (e.g. English). This model is very similar to the concept-based approach adopted for CLIR. Indeed, in the concept-based CLIR the source and target languages are *translated* in a third representation that is generally based on a set of concepts.

Concept-based approaches can adopt implicit and explicit concept models. The most prominent implementations of implicit concept models are Latent Semantic Indexing (LSI) [12] and Latent Dirichlet Allocation (LDA) [6]. Both LSI and LDA perform a

[4] `https://translate.google.com/`
[5] `http://www.microsoft.com/translator/`
[6] `http://www.clef-initiative.eu/edition/clef2009`
[7] `http://www.microsoft.com/translator/`

dimensionality reduction of the document space and the reduced dimensions are the implicit concepts used for indexing new documents. LSI and LDA are exploited for facing CL retrieval tasks [27]. Approaches based on implicit models require a training dataset in order to learn the model. On the other side, explicit concept models exploit concepts whose semantics is explicitly defined. In [47], Sorg et al. propose Cross-Language Explicit Semantic Analysis (CL-ESA) that exploits a semantic representation based on concepts defined by humans. Their approach represents text fragments (e.g. queries, documents) using Wikipedia entries as concepts. The Explicit Semantic Analysis (ESA) technique [18], on which CL-ESA is based, was originally adopted for classification tasks and for computing semantic relatedness between text fragments. The idea underlying CL-ESA is to represent a text fragment in terms of Wikipedia concepts (as ESA already does) and then switching from a language to another by exploiting cross-language links between Wikipedia articles in different languages. Accordingly, given a Wikipedia-based representation, it is straightforward to shift from one language to another. In [8], CL-ESA was compared to approaches based on latent models (LSI and LDA) and showed similar results, even though implicit models need to be trained. The ESA-based representation has also been exploited for a CLIR task characterized by very short documents [36]. In that work the authors combined a translation-based model with a concept-based one. First, all document collections are translated in a pivot language (i.e. English), then the translated text is represented in terms of English Wikipedia concepts. This hybrid model showed better performance than CL-ESA, probably due to the shortness of the available documents, and resulted to be effective on six different European languages, compared to a simpler translation-based model exploiting only keywords.

In [16] a method for solving Cross-Lingual Question Answering based on Wikipedia and EuroWordNet is proposed. The idea is to use several multilingual knowldge sources to reference words between languages without any translation process. The idea is very similar to those proposed in this work, even though it is applied for addressing a different task.

In this article we exploited different adaptations of ESA for facing the cross-lingual recommendation task. We also compared the ESA-based approaches to other approaches based on entity-linking algorithms (e.g. Tagme, Babelfy) based on Wikipedia.

2.3. Multilingual and Cross-Lingual Information Filtering

To the best of our knowledge, very few research on Multilingual and Cross-Lingual IF is available in the literature. An attempt to define an effective multilingual IF system is proposed in [44]. The system is based on the fuzzy set theory. The content of multilingual documents is represented using a set of universal content-based topic profiles, encapsulating all feature variations among multiple languages. Using the co-occurrence statistics of a set of multilingual terms extracted from a parallel corpus (collection of documents containing identical text written in multiple languages), fuzzy clustering is applied to group semantically-related multilingual terms to form topic profiles. The basic intuition is that translated versions of the same text are linguistic variants of the same topic(s), hence, multilingual terms that co-occur in the corresponding translated documents are semantically related to the same topic(s). The main disadvantage of this approach is the need of a significant parallel corpus.

The Multilingual IF task at CLEF 2009[8] has introduced the issues related to the cross-language representation in the area of IF. Damankesh et al. [10] proposed the application of the theory of Human Plausible Reasoning (HPR) to the domain of filtering and cross language IR. The developed system utilizes plausible inferences to infer new unknown knowledge from existing knowledge to retrieve not only documents that match the query terms, but also those which are plausibly relevant.

In [29], the authors proposed an approach to build a model of user interests based on word senses rather than words. The approach relies on MultiWordNet[9] to perform Word Domain Disambiguation and to create synset-based multilingual user profiles, and it has been shown to be effective for news filtering. A similar approach is presented in [28], in which the authors proposed a multilingual content-based recommender system exploiting a semantic representation based on MultiWordNet. Differently from [29], a WSD algorithm has been adopted to obtain a concept-based representation, which showed an accuracy comparable to monolingual suggestions.

In a recent work [43] the authors propose an ontology-based multilingual recommender system using data coming from the Linked Open Data to generate multilingual recommendations in the movie domain. However, the basic idea is to create connections between versions of a given movie on different languages, and not to natively generate cross-lingual recommendations.

Also the adoption of alternative techniques for dimensionality reduction in the areas of monolingual and multilingual IF is relatively new. The use of dimensionality reduction techniques which do not need factorization, such as Random Indexing [25], coupled with the so-called *distributional hypothesis* [23] to build language-independent user profiles, has been investigated in [33]. According to the distributional hypothesis, the meaning of a word is determined by the rules of its usage, i.e. words are semantically similar to the extent that they share contexts (e.g. surrounding words, sentences, documents). The power of distributional approaches is that two terms, in different languages are similar because they share the same context, and this allows to obtain performance of recommendations in a multilingual environment similar to that obtained through WSD [33].

2.4. Recommender Systems

Since this work focuses on a content-based multilingual recommendation process, in the following we provide the basic concepts about the main paradigms to implement recommender systems, and how they deal with the multilingual and cross-lingual recommendation process.

Recommender systems are IF techniques which provide personalized suggestions about items the user might find interesting, by matching items to the user profile. The recommendation problem has been studied extensively, and two main paradigms have emerged:

- *collaborative filtering* [13], which exploits the users' rating style to identify users whose preferences are similar to a given user (neighbors) and recommend items they have liked;

[8] `http://www.clef-campaign.org/2009.html`
[9] `http://multiwordnet.fbk.eu/`

- *content-based filtering* [19], which analyzes a set of documents, usually textual descriptions of items previously rated as relevant by a user, builds a profile of user interests based on the features (usually words) describing the items, and exploits that profile to recommend new relevant items.

In principle, collaborative filtering is inherently cross-lingual, since it does not rely on the content of items for providing recommendations, but solely on the users' rating style, i.e. set of ratings provided by users on items. However, collaborative filtering systems can not be applied in those scenarios where there is a rapid turnover of the recommend items and consequently the new item problem is particularly relevant (e.g. the news recommendation). Indeed, the similarity between users on which collaborative filtering systems are based on, is only computable if they have common rated items. Hence, content-based recommender systems (CBRS) could be adopted, even though traditional CBRS adopt a keyword-based representation for both user profiles and item descriptions. This represents a problem due to the strict connection with the user language: for example, an English user frequently interacts with information written in English, so her profile of interests mainly contains English terms. In order to receive suggestions of items whose textual description is in a different language, she must explicitly give her preferences on items in that specific language. This means that the information already stored in the user profile cannot be exploited to provide suggestions for items whose description is provided in other languages, although they share some common features (i.e. an Italian and an English movie might share the same features, but their plots are written in different languages). This refers to the extreme case of the *vocabulary mismatch* problem described in Section 1. A proposal that do not exploit content neither user ratings and could be exploited in a multilingual scenario is proposed in [30]. However, the approach requires information about the user trust that is not always easy to catch.

In this paper we focus on the use of CBRS leveraging concept-based representations of items and user profiles as a way of providing an effective cross-lingual recommendation process.

3. Cross-lingual Content-based Recommendation

The recommendation process is defined in the literature as the maximization of a utility function that estimates the usefulness (usually expressed by a rating) of an item for a given user [1].

More formally, given a utility function $U : C \times S \rightarrow R$, where C is the set of users, S is the set of items, and R is a totally ordered set, the recommendation process is defined as follows [1]:

$$\forall c \in C, s'_c = argmax_{s \in S}(c, s) \tag{1}$$

Then, for each user $c \in C$, the recommender chooses such item $s'_c \in S$ that maximizes the user utility. In case the system is a monolingual content-based recommender, the user preferences in the profile c are expressed in the same language of the items $s \in S$. Conversely, if the system is a cross-lingual content-based recommender, the problem is more complex than the monolingual case, since there is another argument l to be considered, which represents the language. Let L be the set of languages the system

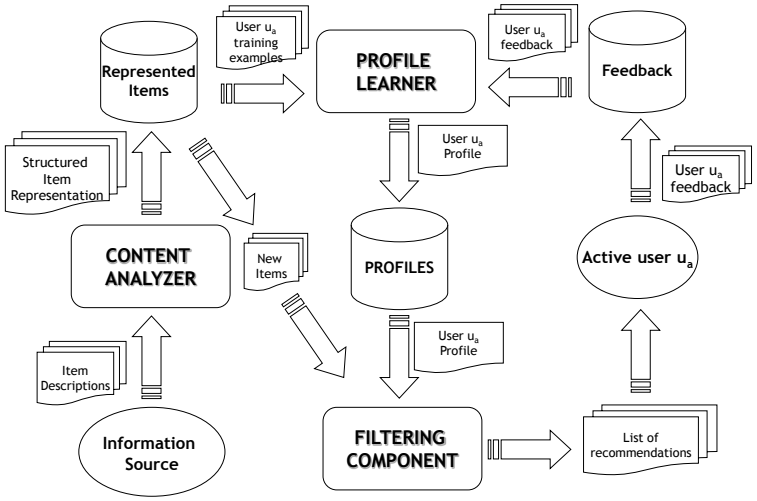


Figure 1: High level architecture of a Content-based Recommender

deals with, let $l_1 \in L$ be the language of the preferences in the user profile, and let l_2 be the language of the recommended items, with $l_1 \neq l_2$, the utility function becomes:

$$\forall c_{l_1} \in C, s'_{c_{l_1}} = argmax_{s_{l_2} \in S}(c_{l_1}, s_{l_2}) \tag{2}$$

Hence, the problem the cross-lingual recommender system has to face is the recommendation of an item $s'_{c_{l_1}}$ in a language l_2 different from the language l_1 of the user profile.

In order to understand which components of a content-based recommender system are involved in the cross-lingual recommendation process, in Figure 1 the high level architecture of a content-based recommender system is reported [19]. The recommendation process is performed as follows: the CONTENT ANALYZER represents items (e.g. product descriptions) in a structured form using specific features (keywords, n-grams, concepts, etc); the PROFILE LEARNER collects data representative of the user preferences in order to automatically build a profile of the user interests; the FILTERING COMPONENT exploits the user profile to suggest relevant items by matching the profile representation against that of items to be recommended.

We address the problem of learning user profiles using Machine Learning techniques [19]. A set of items $S = \{s_1, \ldots, s_n\}$ is labeled by a specific user with relevance judgments (binary or in a discrete scale) that indicate her *degree of interest* (i.e. the utility) in those items. Each item s_i, represented by a set of features and coupled with its relevance judgment, is treated as a single datapoint, and a set of datapoints can be used for training purposes. This allows to learn a function to *predict the relevance judgment* of new unknown items, namely the utility function.

More formally, let $x_i = \phi(s_i)$, where ϕ is a feature extractor and x_i is a m-dimensional vector. Let $TR = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a set of item representations and their associated relevance ratings $y_i \in Y$. In our recommendation scenario, relevance is **1** for the items interesting to that user and **0** for all the other items. TR is used to train a classification model.

It is evident the *language-dependent* nature of such a process when using keywords as features to represent items, which leads to the induction of user profiles in terms of those features. As a consequence, the recommender is only able to suggest other items in which those specific features *explicitly* occur, and this does not allow multi- and cross-lingual recommendations. We can now declare that the only component impacted by the cross-lingual extension of a content-based recommender system is the Content Analyzer. Hence, the item representation plays a crucial role in a cross-lingual content-based recommendation scenario. Therefore, in the next sections we focus our attention on different content representation exploitable for facing the cross-lingual content-based recommendation process. To this purpose, in this work we adopted different implementations of ϕ, which allow to represent items using different kinds of features:

- simple keywords;
- concepts extracted from Wikipedia;
- concepts extracted from BabelNet.

Among different classification methods which could be used to learn user profiles, we adopted Random Forests (RF) [7], already proved to be effective in other recommendation scenarios [4]. RF combines different tree predictors built using different samples of the training data (extracted with replacement from the whole training set) and random subsets of the data features. The class of an item is determined by the *majority voting* of the classes returned by the individual trees. The use of different samples of the data from the same distribution and of different sets of features for learning the individual decision trees prevent the overfitting.

The cross-lingual recommendation process can be also viewed as a particular case of *cold start* problem. Indeed we can easily identify the new item problem and the new user problem. The new item problem is quite similar to the monolingual scenario, thus a content-based approach can effectively address it. The situation is more complex when the new item problem is associated to the new user problem. In this case the situation is different from the monolingual scenario, since it includes also the case in which the user profile is partially in cold start, namely it contains only preferences in the language l_1, and the user would receive recommendations in a language l_2, with $l_1 \neq l_2$. For better understand the problem, we propose a typical use case.

Anna is an Italian manager who regularly spends some time in Paris. She loves to watch movies at home or at cinema in her spare time. She uses a platform that recommends movies according to her preferences and past viewings and she is really satisfied by the service provided. Her profile is mostly composed of Italian movies, since she mainly uses the platform in her country. When she is abroad, she likes to go to cinema, but the systems suggests her Italian movies since her user profile contains only Italian features. Hence, according to the previous definition, we can state that the Anna's profile is in cold start for the French language and she could benefit from cross-lingual recommendations. Furthermore, it is not easy to find French versions of the suggested Italian movies since Italian and French editions come out in different time. Therefore, she would like to have a system able to recommend new French movies (*new item problem*) according to her Italian movie preferences. This is a typical use case where a recommender system able to exploit information stored in the user profile in a given

language for recommending new items in another language can be very useful. This scenario can be effectively addressed by the cross-lingual content-based recommender system described in the previous section.

4. Knowledge-based Strategies to build Concept-based Item Representations

The use of external knowledge sources can be useful to better understand the information items (documents, news, product descriptions) and to extract meaningful features in order to have better representations.

Among unstructured knowledge sources, Wikipedia emerges as the most used one for several tasks [14, 24], since it is free and covers many domains, it is very accurate [20] and available in several languages. On the other hand, Wikipedia content is available in textual form written by humans for humans, and needs to be processed for becoming *machine processable*. The problem of extracting and using knowledge contained in Wikipedia has been studied by several researchers [9, 18, 15]. Several techniques have been defined, which exploit the encyclopedic knowledge contained in Wikipedia for *selecting the most accurate semantic features* to represent the items, or for *generating new semantic features* to enrich the item representation.

In this work we compare the effectiveness of four distinct knowledge-based strategies which exploit different knowledge sources to build concept-based representations, in order to provide cross-lingual recommendations:

- *Tagme* – an entity linking algorithm based on Wikipedia (Section 4.1);
- *Explicit Semantic Analysis* (ESA) – a method leveraging the unstructured encyclopedic knowledge contained in Wikipedia (Section 4.2);
- *Babelfy* – an integrated approach to entity linking and Word Sense Disambiguation based on BabelNet (Section 4.3.1);
- *Distributional Lesk-Word Sense Disambiguation and Entity Linking* (DL-WSDEL) – a combination of entity linking and Word Sense Disambiguation based on BabelNet (Section 4.3.2).

Table 1: Plot summary for the movie "Rocky"

Title	Rocky
Source	`http://en.wikipedia.org/wiki/Rocky`
Plot Summary	Rocky is a 1976 American sports drama film directed by John G. Avildsen and both written by and starring Sylvester Stallone. It tells the rags to riches American Dream story of Rocky Balboa, an uneducated but kind-hearted debt collector for a loan shark in the city of Philadelphia, Pennsylvania. Rocky starts out as a club fighter who later gets a shot at the world heavyweight championship.

It is worth to note that Babelfy and DL-WSDEL are the only methods based on the use of a *native* multilingual knowledge source, i.e. BabelNet, differently from the

other two methods which are based on Wikipedia, and for which a specific processing to deal with multilinguality is needed. Another difference is that, differently from the other techniques, ESA is the only one able to *generate* new semantic features to enrich item representations [18].

In order to better explain the different concept-based item representations evaluated in this work, we will use the plot summary of the movie "Rocky" in Table 1 as a running example throughout the paper, and we will show how that plot summary is represented using the above mentioned knowledge sources and techniques.

4.1. Concept-based Representation based on Tagme

Tagme [15] is an *entity linking* algorithm able to produce a rich and fine-grained semantic content representation relying on Wikipedia-based features. Entity Linking (EL) [45] techniques aim to map an input text (typically tokenized in n words, $w_1 \dots w_n$) to k entities ($e_1 \dots e_k$, $k \leq n$) that are mentioned in it.

Tagme adopts Wikipedia as knowledge base, leading to a broad coverage of the concepts that can be potentially linked. The *linking methodology* is carried out in three steps: 1) *anchor parsing* – to scan the input text to identify all the potential mentions to entities; 2) *anchor disambiguation* – to identify the correct entity (i.e. the Wikipedia page) the anchor actually refers to; and 3) *anchor pruning* – once the disambiguation is performed, the final set of anchors is pruned in order to filter out noisy mentions. The output of the process is a set of entities each of which is provided with a confidence score. The Tagme representation of the plot summary in Table 1 is reported in Table 2.

Table 2: Tagme-based representation of the plot summary for the movie "Rocky".

Tagme concepts	**URL** `https://en.wikipedia.org`
Drama film	`/wiki/Drama_film`
Film director	`/wiki/Film_director`
John G. Avildsen	`/wiki/John_G._Avildsen`
Sylvester Stallone	`/wiki/Sylvester_Stallone`
American Dream	`/wiki/American_Dream`
Rocky Balboa	`/wiki/Rocky_Balboa`
Philadelphia	`/wiki/Philadelphia`
Club fighter	`/wiki/Club_fighter`
World Heavyweight Championship (WWE)	`/wiki/World_Heavyweight_Championship_(WWE)`

Concept-based representations built exploiting EL algorithms as Tagme could lead to a language independent representation of items. This is possible since the suffixes used in Wikipedia pages to refer to named entities (e.g. cities, actors, directors, ...) in different languages are usually the same. For example the English and Italian Wikipedia pages for *Sylvester Stallone* are `https://en.wikipedia.org/wiki/Sylvester_Stallone` and `https://it.wikipedia.org/wiki/Sylvester_Stallone`, respectively, which allow to match the concept *Sylvester Stallone* in different languages. Another possible way to match concepts in different languages is to leverage the *cross-language* links between Wikipedia articles, which allow to refer to the same Wikipedia article (concept) in different languages. The main limitation of Tagme is the availability only in English and

Italian. Hence, dealing with other languages requires a translation process from a language not supported by Tagme into English or Italian.

4.2. Concept-based Representation based on Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) [18] is an approach which leverages Wikipedia to *generate* new features for enriching item representation. ESA provides a fine-grained semantic representation of text documents as a weighted vector of concepts derived from Wikipedia. Specifically, concepts correspond to Wikipedia articles, such as Woody Allen, or Apple Inc.. ESA resembles the well known LSA technique [12], whose representation is based on *latent* (and not comprehensible) features, rather than *explicit* (and comprehensible) concepts derived from Wikipedia (concepts explicitly defined and manipulated by humans).

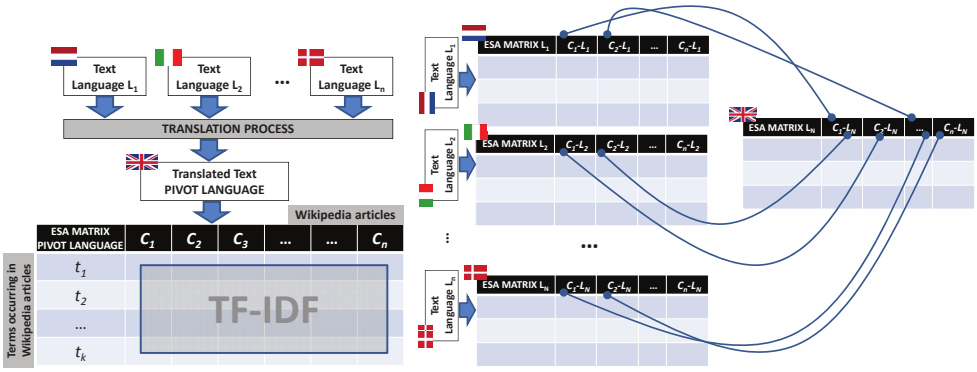
The idea behind ESA is to view an encyclopedia as a collection of concepts, each of which accompanied with a large body of text (the article content). The power of ESA is the capability of representing the Wikipedia knowledge base in a way that is directly used by machines. The gist of the technique is to use the high-dimensional space defined by these concepts in order to represent the meaning of natural language texts. ESA allows to leverage Wikipedia knowledge by defining relationships between terms and Wikipedia articles. More formally, given a set of basic concepts $C = \{c_1, c_2, ..., c_n\}$, a term t is represented by a vector of weights $< w_1, w_2, ..., w_n >$, where w_i represents the strength of association between t and c_i. The set of concepts C are one by one associated to documents $D = \{d_1, d_2, ..., d_n\}$ (the Wikipedia articles). Hence, a sparse matrix T is built, called *ESA-matrix*, where each column corresponds to a concept (title of Wikipedia article), and each row corresponds to a term that occurs in D. The entry $T[i, j]$ of the matrix represents the TF-IDF of term t_i in document d_j. Finally, length normalization is applied to each column to disregard differences in document length. This allows to define the semantics of a term t_i as a point in the n-dimensional semantic space of Wikipedia concepts. The weighted vector corresponding to a term t_i is called *semantic interpretation vector*. The semantics of a text fragment $< t_1, t_2, ..., t_k >$ (i.e. a sentence, a paragraph, an entire document) is obtained by computing the centroid (average vector) of the semantic interpretation vectors of the individual terms occurring in the fragment.

The ESA representation of the plot summary in Table 1 is reported in Table 3. It

Table 3: ESA-based representation of the plot summary for the movie "Rocky".

ESA concepts	**Description**
World Heavyweight Champions	The list of heavyweight boxing champions
David Bey	A former USBA heavyweight champion who challenged the legendary Larry Holmes for the world title in 1985
Jack London	Real name John George Harper, an English heavyweight boxer
Doug Jones	A former American heavyweight boxer
Fight Club (video game)	A fighting video game based on the film Fight Club
Super heavyweight	In amateur boxing, the super heavyweight division is a a weight class division for fighters weighing in excess of 91 kilograms
My Love (Celine Dion song)	The lead single from Céline Dion's greatest hits album My Love: Essential Collection

is worth to notice that ESA is able to generate new knowledge in terms of Wikipedia concepts which do not directly occur in the plot summary of the movie "Rocky" (Table



(a) Translation-based ESA (b) Cross-language ESA

Figure 2: Cross-lingual document representation using ESA

1). At first sight, some concepts seem to be not related to the movie topic (My Love), but those who have seen Rocky know that the movie narrates the love story between Rocky and Adrian as well.

Two different ways of using ESA to obtain a language independent document representation are possible: translation-based ESA (TR-ESA), and cross-language ESA (CL-ESA) [8].

In TR-ESA, (Figure 2a) documents in different languages are represented using Wikipedia concepts in a given unique language, called *pivot language*. This allows to provide a common representation of documents in different languages in the same space of Wikipedia concepts, leading to documents which are thus directly comparable. Texts in different languages are first translated in the pivot language (e.g. English) and then semantically-represented using the ESA matrix corresponding to the pivot language.

Conversely, CL-ESA needs an ESA matrix for each language in which documents are represented (Figure 2b). In order to make the concept-based representations directly comparable, it is possible to leverage the *cross-language links* between Wikipedia articles in different languages. For example, given a Dutch and an Italian document, the former is represented in terms of Dutch Wikipedia concepts (through the Dutch ESA matrix), and the latter is represented in terms of Italian Wikipedia concepts (through the Italian ESA matrix). We could create a direct link between Dutch and Italian Wikipedia concepts using the cross-language links, or we could create a link of Dutch and Italian Wikipedia concepts towards concepts of an ESA matrix corresponding to another language, e.g. English.

4.3. Concept-based Representation based on BabelNet

In this section we introduce *BabelNet*[10] [38], a knowledge resource that offers a multilingual coverage of both lexicographic senses and encyclopedic information by integrating

[10] `http://babelnet.org`

Wikipedia and WordNet[11]. We also present two strategies based on BabelNet to build concept-based representations, which could be used to provide cross-lingual recommendations:

- *Babelfy*[12], a novel unified graph-based approach to entity linking and Word Sense Disambiguation;
- *DL-WSDEL*[3] which combines WSD based on BabelNet, with a specific model for named entity discovery.

BabelNet encodes knowledge as a labeled directed graph: *nodes* are *concepts* extracted from WordNet and Wikipedia, i.e. word senses (synsets) available in WordNet, and encyclopedic entries (Wikipages) extracted from Wikipedia, while *edges* connecting the nodes are labeled with *semantic relations* coming from WordNet, as well as semantically unspecified relations from hyperlinked text coming from Wikipedia. The resulting representation in BabelNet is a set of *Babel synsets* connected through semantic relations. The first synset comes from Wikipedia, while the second one was collected from WordNet and belongs to the synonym set {rocky, bouldery, bouldered, stony}. Each concept is natively associated with a set of lexicalizations in the different languages. For example, the first sense of the adjective "rocky" is associated to the Spanish "pedregoso, rocoso", the German "steinig, felsig", the Italian "sassoso, petroso, roccioso", and the Russian "каменистый, скалистый". Moreover, each concept is provided with one or more glosses, preferably in different languages. The gloss, which gives an explanation of the sense, can be regarded as a defining context for that sense.

The current version of BabelNet (3.6) covers 272 languages, and contains 13.8 millions Babel synsets and 380 millions of lexico-semantic relations.

4.3.1. Babelfy

Babelfy is a novel, unified graph-based approach to EL, whose output is a bag of Babel synsets, each identifying in a unique way concepts and named entities in different languages. Babelfy uses BabelNet 1.1.1 [38] and the details of the techniques are reported in [32]. An excerpt of the Babelfy synset-based representation of the plot summary in Table 1 is reported in Table 4.

The main advantage of Babelfy is the unified approach on the two tasks of EL and WSD in any of the languages covered by the *native* multilingual semantic network.

4.3.2. Distributional Lesk-Word Sense Disambiguation algorithm

As alternative to Babelfy, we adopted DL-WSDEL [3], an extension of the Distributional Lesk-Word Sense Disambiguation algorithm (DL-WSD) [2] that combines WSD based on BabelNet synsets with a specific model for named entity discovery.

In order to produce a Babel synset-based representation of textual content, we need to distinguish entity mentions from broad concepts (e.g. "Rocky movie" from "stony meaning"), and generally both named entities and words need to be disambiguated (rocky as noun has twelve different meanings, while the adjective form is associated with four

[11] https://wordnet.princeton.edu/
[12] `http://babelfy.org`

Table 4: Babelfy representation of the plot summary for the movie "Rocky".

Babel synsets	**Meaning**	**Type**	**Glosses (in English)**
bn:01564901n	American sports	Named Entity	Sports are an important part of the culture of the US
bn:03688355n	drama film	Concept	Film genre that depends mostly on in-depth development of realistic characters dealing with emotional themes
bn:00086865v	directed	Concept	Command with authority
bn:03698290n	John G.	Named Entity	John Guilbert Avildsen is an American film director
bn:00085489v	written	Concept	Produce a literary work
bn:00094295v	starring	Concept	Be the star in a performance
bn:03449858n	Sylvester	Named Entity	Sylvester Gardenzio Stallone, nicknamed Sly Stallone, is an American actor, screenwriter and film director
...			

different concepts). DL-WSDEL algorithm consists of two steps: 1) *entity recognition* – identification of all possible entities mentioned in a text, associated with a set of possible meanings, i.e. Babel synsets; 2) *WSD* – disambiguation of both words and named entities through the DL-WSD algorithm.

Words and named entities are disambiguated using the distributional Lesk algorithm [2], which replaces the concept of *word overlap*, initially introduced by Lesk [26], with the broader concept of *semantic similarity*. The novelty of the approach is that the similarity is computed by representing both the gloss and the context in a Distributional Semantic Model (DSM). DSMs rely on the *distributional hypothesis* [23], according to which "Words that occur in the same contexts tend to have similar meanings". Hence, the semantic representation of terms is directly learned according to the way terms are used in a large corpus of data. In this work we have built a DSM by analyzing the whole set of pages in the Wikipedia dump of the language for which we want to build the representation. A word/entity is disambiguated by choosing the sense whose gloss maximizes the semantic similarity with the word/entity context. The presentation of the details of the algorithm is out of the scope of the current work and can be found in [3]. The output of DL-WSDEL is a bag of Babel synsets, each identifying in a unique way concepts and named entities in different languages. We exploited BabelNet 2.5.1 that covers 50 languages.

The Babel synsets-based representation of the plot summary in Table 1 is reported in Table 5.

Table 5: Babel synset-based representation of the plot summary for the movie "Rocky".

Babel synsets	**Meaning**	**Type**	**Glosses (in English)**
bn:03220034n	Rocky	Named Entity	Rocky is a 1976 American sports drama film directed by John G. Avildsen and both written by and starring Sylvester Stallone...
bn:00096963a	American	Concept	Of, from, or pertaining to the USA, its people or its culture...
bn:00006759n	sport	Concept	Competitive physical activity...
bn:00034471n	movie	Concept	A film, movie.
...	...	...	...

The main advantage of this approach is the use of a *native* multilingual knowledge source, while the main limitation is the need of defining a DSM based on Wikipedia for each language to deal with.

5. Experimental evaluation

The main goal of the experimental evaluation is to compare the effectiveness of the different knowledge-based strategies described in Section 4, in order to validate the hypothesis that a concept-based representation of items and user profiles is an effective strategy to provide cross-lingual recommendations. More specifically, we would like to test:

1. the performance of the different concept-based representations to produce cross-lingual recommendations;
2. the influence of the length of the item descriptions on the accuracy of recommendations;
3. the influence of the translation direction (pivot language) on the accuracy of recommendations.

Experiments are carried out on two datasets, i.e. `DBbook` and `MovieLens`, and are focused on two languages, namely English and Italian.

5.1. Datasets and Evaluation Measures

The `DBbook` dataset[13] comes from the Linked-Open Data-enabled Recommender Systems challenge and focuses on book recommendations, while `MovieLens`[14] is a widespread dataset for movie recommendations.

From the original `DBbook` dataset we kept only those items for which a Wikipedia page in both English and Italian was available. Hence, we filtered out those users who provided less than 5 ratings. We obtained a subset of the `DBbook` dataset containing $2,362$ items, rated by $5,095$ users, who provided $74,048$ ratings (sparsity 99.38%). For `MovieLens` the same process leads to a dataset with $90,096$ ratings, provided by 943 users on $1,235$ items (sparsity 92.26%)[15]. We decided to take into account only binary ratings. `DBbook` is already provided with binary ratings, thus no further processing was needed. On the other side, given that `MovieLens` preferences are expressed on a 5-point Likert scale, we deemed as *positive* those ratings equal to 4 and 5.

Some statistics about the datasets are provided in Table 6.

`DBbook` is more sparse than `MovieLens`, along with an average number of ratings per user significantly lower than that provided by `MovieLens` users (14.53 vs. 95.54 ratings on average, 10 vs. 20 as mode). This makes `MovieLens` more suitable for learning more accurate content-based user profiles. Both the datasets have a similar balance in terms of positive and negative ratings.

[13] http://challenges.2014.eswc-conferences.org/index.php/RecSys#DBbook_dataset

[14] http://grouplens.org/datasets/movielens/

[15] The datasets are available for download at the following link:
`https://www.dropbox.com/sh/u9pj75y2sw3d4gi/AACv9CMQttQ7PVwFtIY5ArtPa?dl=0`

Table 6: Statistics about the datasets.

	DBbook	MovieLens
Users/Items	5,095 / 2,362	943 / 1,235
Ratings (sparsity)	74,048 (99.38%)	90,096 (92.26%)
Positive Ratings	57.11%	56.40%
Ratings per user: mean	14.53 (σ=5.23)	95.54 (σ=91.16)
Ratings per user: median/mode	14/10	60/20

In order to evaluate the influence of the length of the item descriptions on the accuracy of recommendations, we considered the whole Wikipedia page describing items, or just the abstract, i.e. the first paragraph of the Wikipedia page. Table 7 presents the statistics about the average length of the item descriptions, for both English and Italian. As regards Tagme, we only kept concepts with a confidence score greater or equal than 0.05, while for ESA, we adopted *multiresolution* by taking into account sentences as segments, and by considering *all* the concepts generated by ESA for abstracts, and the *10-most-related* concepts for Wikipedia descriptions.

Table 7: Average number of features to represent item descriptions.

Language	**Represent.**	DBbook		MovieLens	
		Abstract	Wikipedia	Abstract	Wikipedia
ITA	Keywords	43	266	31	247
	Tagme	-58%	-67%	-58%	-67%
	Cl-Esa	>+1,000%	+31%	>+1,000%	+46%
	TR-Esa	>+1,000%	+42%	>+1,000%	+52%
	Dl-Wsdel	0%	-8%	0%	-8%
	Babelfy	-63%	-75%	-55%	-76%
ENG	Keywords	59	546	67	581
	Tagme	-58%	-60%	-55%	-66%
	Cl-Esa	>+1,000%	+31%	>+1,000%	+46%
	TR-Esa	>+1,000%	+42%	>+1,000%	+52%
	Dl-Wsdel	-8%	-6%	-4%	-4%
	Babelfy	-54%	-70%	-48%	-70%

The average length of item descriptions in the two datasets is comparable, and English descriptions are longer than Italian ones, especially for Wikipedia. Tagme and Babelfy represent items using a similar number of features, while DL-WSDEL adopts a number of features which is very similar to the keyword-based representation. It is also evident that, even though DL-WSDEL and Babelfy are both based on BabelNet, they use a very different number of features to represent item descriptions, with Babelfy adopting a considerably smaller number of features than DL-WSDEL. As expected, ESA is the only method which increases the number of features with respect to the original number of keywords, due to the enrichment process which generates new semantic features based on Wikipedia. This is particularly evident when all the features are taken into account, as in the representation of abstracts.

As our content-based recommender system is conceived as a classifier able to discriminate items as interesting or not for each specific user, its effectiveness is evaluated using F1 measure, the harmonic mean of precision and recall, where *precision* is the ratio between the number of correctly classified items and the number of classified items, and *recall* is the ratio between the number of correctly classified items and the total number of classified items. For the sake of brevity, Section 5.3 only reports F1 figures (average values of like and dislike classification).

5.2. Design of the Experiments

Experiments were carried out using a *per user* evaluation, organized as follows:

1. ratings of the active user u_a (for which recommendations must be provided) are split into a training set Tr and a test set Ts, using a 70%-30% training-test split;
2. ratings in Tr along with the corresponding item descriptions are used to learn the user profile of u_a using Random Forests. The user profile is learned using several configuration obtained by varying the item representations strategies, the length of item descriptions (abstract or whole Wikipedia page), and the pivot language (Italian or English), which is the language chosen to have a common representation of item descriptions in Tr and Ts. We performed experiments in which item descriptions in Tr are in Italian, while those in Ts are in English (ITA-ENG). This corresponds to learn a content-based user profile from items whose description is in Italian, and providing recommendations on items whose description is in English. Similarly, we performed experiments in which item descriptions in Tr are in English, while item descriptions in Ts are in Italian (ENG-ITA);
3. the predictive accuracy of the user profile of u_a is computed on items in Ts;
4. results are averaged for all users.

Profiles are classifiers learned using Random Forests. We adopted the Weka[16] implementation using the default parameters.

The use of different strategies to represent item descriptions allows to evaluate their ability to really provide effective cross-lingual recommendations. The use of item representations coming from abstracts or the whole Wikipedia pages is useful to assess the influence of the description length on the overall accuracy of the recommendations. The influence of the pivot language on the accuracy of recommendations is evaluated by taking into account the representation or the translation of the description of items occurring in the training set or in the test set.

To summarize, for each dataset, we evaluate the following approaches, for which we report the different configurations:

- *Keyword-based approaches*: this is the *baseline*, which adopts a representation of item descriptions based only on *keywords*. The vocabulary mismatch due to the multilingual setting is dealt through a simple translation process of the item descriptions in Tr or Ts;

[16] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/RandomForest.html

- *Concept-based approaches*: do not exploit any translation process, rather they use concepts as a way for providing a language-independent representation, which allows to deal with the vocabulary mismatch problem. We used *Tagme* and *CL-ESA* techniques which represent items using concepts extracted from *Wikipedia*, and *DL-WSDEL* and *Babelfy*, which leverage *BabelNet*.

- *Translation-based approaches*: use concept-based representations as in the previous approach, with the only difference that the content is previously translated in the pivot language. This means that item descriptions in *Tr* and *Ts* are first translated in the same language, and then represented using concepts.

The translation process is performed using Bing translator[17].

The *t-test* has been performed to assess statistically significant differences between F1 values ($p < 0.05$) of the different configurations. For the sake of brevity, we only report results obtained with the best pivot language, namely the same language of the training set. More details in the next Section.

5.3. Discussion of Results

Figures 3 and 4 report the results for `DBbook`, while Figures 5 and 6 report the results for `MovieLens`. We will analyze the results from different perspectives.

As expected, recommendations obtained on `MovieLens` are more accurate than those obtained on `DBbook`. Indeed, the lower sparsity of `MovieLens` with respect to `DBbook`, and the higher average number of ratings per user lead to better content-based user profiles.

As regards the influence of the length of the item descriptions on the accuracy of recommendations, it is worth to note that, most of the times the approaches based on the abstract outperform those based on Wikipedia, regardless the adopted representation, i.e. keyword-based, concept-based, or translation-based. This is an advantage since it is not necessary to process very long item descriptions to obtain better results.

As regards `DBbook`, the best *concept-based approach* is the one based on Babelfy, with item descriptions extracted from the abstract, regardless the use of English or Italian as language for representing the training set, while the best *translation-based approach* is still based on the abstract, and corresponds to TR-ESA. The latter is also the best performing approach for the ITA-ENG experiment. The differences between TR-ESA and the best concept-based approach (i.e. Babelfy) and the best keyword-based configuration are statistically significant. For ENG-ITA the best overall approach is based on keywords and Wikipedia as source for descriptions. However, ENG-ITA experiment confirms ITA-ENG results, namely that best translation-based approach is TR-ESA and the best concept-based approach is Babelfy. Also in this case the differences between the best overall approach (i.e. the keyword-based one) and the best concept-based (i.e. Babelfy) and translation-based (i.e. TR-ESA) approaches are statistically significant. The advantage of having Babelfy as the best concept-based approach is that it is based on BabelNet, a native multilingual knowledge repository, there is no need of any translation process, and the bootstrap for a new language is very simple. The bootstrap for a new language with the TR-ESA technique is quite simple as well, since a unique ESA matrix has to been built.

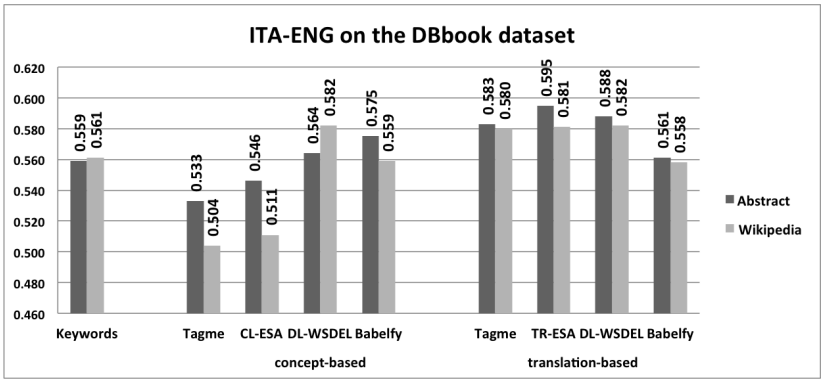


Figure 3: F-measure for ITA-ENG on the DBbook dataset

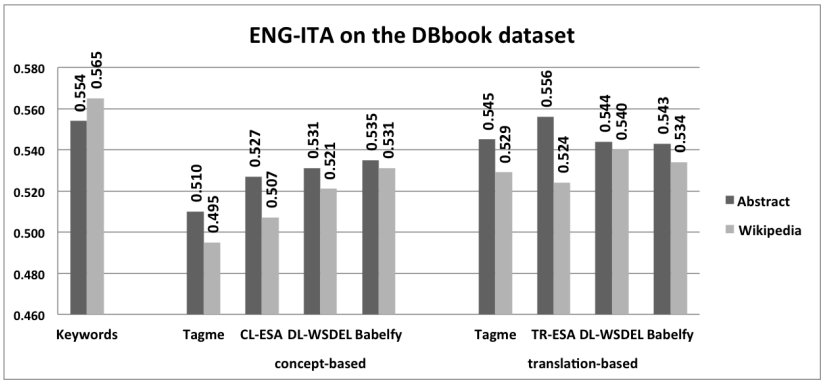


Figure 4: F-measure for ENG-ITA on the DBbook dataset

As regards `MovieLens`, the best *concept-based approach* is DL-WSDEL for ITA-ENG, where the source for descriptions is Wikipedia, while CL-ESA working on the abstract for ENG-ITA. DL-WSDEL has the advantage of using BabelNet, but the bootstrap of a new language is not simple, since it needs to build a Distributional Semantic Model based on Wikipedia in that language. On the other side, CL-ESA requires a different ESA matrix for each language it deals with, and the alignment of concepts in the different languages must be provided. The best *translation-based approach* is TR-ESA based on the abstract for ITA-ENG, and DL-WSDEL based on Wikipedia for ENG-ITA. As for `DBbook`, the best performing approach for `MovieLens` is TR-ESA based on the abstract for ITA-ENG. The difference between TR-ESA and the best keyword-based configuration is statistically significant, while the difference with respect to DL-WSDEL is not statistically significant. For ENG-ITA a keyword-based representation (based on the abstract) is the best performing approach and the differences with respect the best concept-based approach (i.e. CL-ESA) and the best translation-based approach (i.e. DL-WSDEL) are statistically significant. A possible interpretation of the high performance of keyword-

[17] `http://www.bing.com/translator/`

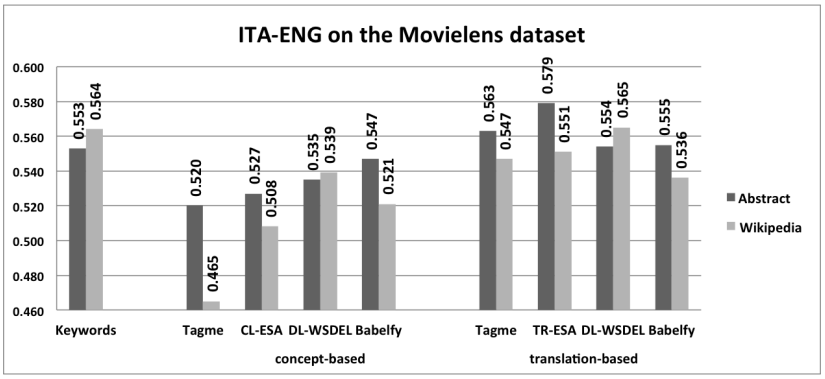


Figure 5: F-measure for ITA-ENG on the Movielens dataset

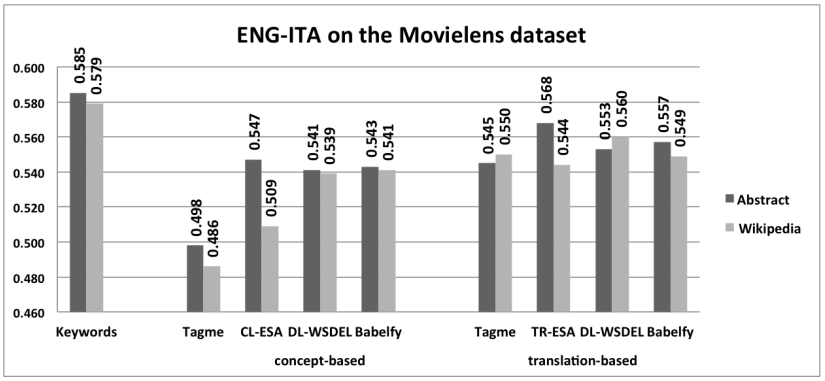


Figure 6: F-measure for ENG-ITA on the Movielens dataset

based approaches for the ENG-ITA experiments for both the datasets is the likely high accuracy of the translation process starting from the Italian language (both experiments have English as pivot language).

Even though the results of the experiments do not show any clear advantage of using one of the proposed representations, it is evident that concept-based approaches which abstract from surface representations are suitable for cross-lingual scenarios. Among the techniques proposed in Section 4, DL-WSDEL and Babelfy generally outperform Tagme and CL-ESA, which led us to conclude that knowledge-based strategies leveraging a native multilingual knowledge source, such as BabelNet, are the most suitable for building language independent item representations. Babelfy has also the additional advantage of using a very compact representation in terms of number of concepts. Unexpectedly, the larger number of (Wikipedia) concepts adopted by CL-ESA-based representations does not lead to better results, especially when compared to Babelfy, which adopts representations based on Wikipedia concepts as well. It is worth noting that TR-ESA on the abstract turned out to be the best approach on three out of four experiments carried out. This is a very interesting result since this approach is exploitable in those scenarios on which items are provided with limited textual content, outperforming keyword-based

approaches at a limited cost.

We also investigated the impact of the pivot language on the recommendation accuracy. For the sake of brevity, we do not report detailed results of this experiment, but we just summarize the main outcomes. The pivot language is the language on which both the training set and the test set are evenly represented before any further processing. For example, in the experiment ITA-EN, if the pivot language is ITA, the test set (EN) is translated in Italian, conversely, if the pivot language is EN, the training set (IT) is translated in English. We adopted as pivot language both English and Italian. The first outcome is that, for both the datasets, there is a quite strong influence of the pivot language on the accuracy of recommendations for keyword-based approaches. More specifically, for ITA-ENG experiments, the best results are obtained using Italian as pivot language, while for ENG-ITA experiments, the best results are obtained using English as pivot language. This result could be interpreted in different ways: on one hand there might be a different performance of the Bing translator on the two different languages, while on the other hand there might be differences due to the translation of the item descriptions contained in the training or test set. Indeed, we noticed that most of the times the approaches based on the translation of the test set outperform those based on the translation of the training set, and this is probably due to the smaller number of translated item descriptions, hence to a smaller number of translation mistakes. The influence of the pivot language is very limited for translation-based approaches, even though, as for keyword-based approaches, most of the times the translation of item descriptions in the test set returns better results than the translation of item descriptions in the training set. Of course, the pivot language does not have any influence for concept-based approaches, since there is not a translation process to match item descriptions in different languages.

Now we can answer to the research questions formulated in Section 1.

- **R1**: Are knowledge-based strategies able to face the content-based cross-lingual recommendation problem?
  Yes. We showed how to adopt different knowledge-based strategies to build concept-based representations of items. Results demonstrated the effectiveness of these strategies in terms of recommendation accuracy and generally they outperform the keyword-based baseline.

- **R2**: Are concept-based representations able to provide effective content-based cross-lingual recommendations compared to translation-based approaches?
  Partially. We showed that concept-based approach, which do not integrate any translation process, generally have slightly worse performance than the translation-based approaches. However, they represent an effective solution to have a native language-independent representation.

- **R3**: Are knowledge-based representations effective to provide content-based cross-lingual recommendations when limited textual content is available?
  Yes. We showed that knowledge-based representations work well and better than the baseline on short textual content. Hence, the knowledge-based strategies are useful when limited content is available, as well.

6. Conclusions and future work

The growing of multilingual content and the increasing number of polyglot users call for new platforms for intelligent information access able to deal with multilingual information. Recommender systems may also benefit from multilinguality, to allow the suggestion of content in languages different from that the user adopted to express his/her interests. In this paper we proposed a strategy to cope with the problem of providing cross-lingual content-based recommendations. The main intuition behind our work is that, a keyword-based representation of content is inherently language-dependent – keywords are different in each language – while concept-based representations are inherently language independent – the meaning of a keyword is the same in different languages.

We proposed four concept-based representations built by exploiting two wide-coverage knowledge sources, namely Wikipedia and BabelNet, and we performed an experimental evaluation on two state-of-the-art datasets to show the effectiveness of the cross-lingual recommendation process. We compared pure concept-based representations with translation-based ones which perform a preliminary translation process before the concept generation.

Preliminary results confirm our hypotheses, namely that:

- knowledge-based strategies are able to face the content-based cross-lingual recommendation problem;
- the concept-based representations can provide quite effective content-based cross-lingual recommendations;
- knowledge-based representations are effective to provide content-based cross-lingual representation when limited textual content is available.

Future work regards the evaluation on different languages and different approaches to come up with concept-based representations. As regards the former aspect, we plan to investigate the effectiveness of the proposed approaches on other languages besides English and Italian. Most of the knowledge-based strategies described in Section 4 have been already adopted to deal with different languages, even though not in an information filtering scenario. Indeed, Tagme and TR-ESA have been adopted in [36] to develop a cross-language e-gov service retrieval system whose catalogs are in Dutch, Belgian, German, Swedish and Norwegian; ESA was adopted in [34] to provide an enhanced semantic representation of German TV-shows descriptions, with the aim of retrieving the most related shows for a specific program type; BabelNet was used in the SemEval-2013 task on Multilingual Word Sense Disambiguation [37], for content available in English, French, German and Spanish.

As for the latter aspect, we are planning to evaluate alternative strategies to provide concept-based representations, which are not based on the adoption of *exogenous* knowledge sources anymore, rather on the use of *endogenous* knowledge coming from the analysis of the rules of usage of terms. These approaches are based on the so-called, *distributional hypothesis* [23], i.e. the meaning of a word is determined by the rules of its usage, that is the co-occurrence with other terms. It is assumed that in every language each term often co-occurs with the same other terms (expressed in different languages, of course), thus by representing content-based user profiles in terms of the co-occurrences

of its terms, user preferences could become inherently independent from the language and this could be sufficient to provide the user with cross-language recommendations. A preliminary investigation carried out in [33] already shows the effectiveness of the approach.

References

[1] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge and Data Engineering 17 (2005) 734–749.

[2] P. Basile, A. Caputo, G. Semeraro, An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 1591–1600.

[3] P. Basile, A. Caputo, G. Semeraro, UNIBA: Combining Distributional Semantic Models and Sense Distribution for Multilingual All-Words Sense Disambiguation and Entity Linking, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 360–364.

[4] P. Basile, C. Musto, M. de Gemmis, P. Lops, F. Narducci, G. Semeraro, Content-Based Recommender Systems+DBpedia Knowledge=Semantics-Aware Recommender Systems, in: V. Presutti, M. Stankovic, E. Cambria, I. Cantador, A.D. Iorio, T.D. Noia, C. Lange, D.R. Recupero, A. Tordai (Eds.), Semantic Web Evaluation Challenge at ESWC 2014, Revised Selected Papers, volume 475 of *Communications in Computer and Information Science*, Springer, 2014, pp. 163–169.

[5] N.J. Belkin, W.B. Croft, Information filtering and information retrieval: Two sides of the same coin?, Communications of the ACM 35 (1992) 29–38.

[6] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet Allocation, the Journal of Machine Learning Research 3 (2003) 993–1022.

[7] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[8] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, S. Staab, Explicit Versus Latent Concept Models for Cross-Language Information Retrieval, in: IJCAI 2009, volume 9, pp. 1513–1518.

[9] A. Csomai, R. Mihalcea, Linking Documents to Encyclopedic Knowledge, IEEE Intelligent Systems 23 (2008) 34–41. doi:10.1109/MIS.2008.86.

[10] A. Damankesh, J. Singh, F. Jahedpari, K. Shaalan, F. Oroumchian, Using Human Plausible Reasoning as a Framework for Multilingual Information Filtering, in: CLEF 2009: Proc. of the 9th Workshop of the Cross-Language Evaluation Forum.

[11] M.W. Davis, T.E. Dunning, A TREC Evaluation of Query Translation Methods for Multilingual Text Retrieval, in: Proceedings of TREC Conference 1995, pp. 483–497.

[12] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, Indexing by Latent Semantic Analysis, JASIS 41 (1990) 391–407.

[13] C. Desrosiers, G. Karypis, A Comprehensive Survey of Neighborhood-based Recommendation Methods, in: F. Ricci, L. Rokach, B. Shapira, P.B. Kantor (Eds.), Recommender Systems Handbook, Springer, 2011, pp. 107–144.

[14] O. Egozi, E. Gabrilovich, S. Markovitch, Concept-based Feature Generation and Selection for Information Retrieval, in: Proc. of the 23rd National Conference on Artificial Intelligence - Volume 2, AAAI'08, AAAI Press, 2008, pp. 1132–1137.

[15] P. Ferragina, U. Scaiella, Fast and Accurate Annotation of Short Texts with Wikipedia Pages, IEEE Software 29 (2012) 70–75.

[16] S. Ferrández, A. Toral, Óscar Ferrández, A. Ferrández, R. Muñoz, Exploiting Wikipedia and EuroWordNet to solve Cross-Lingual Question Answering, Information Sciences 179 (2009) 3473 – 3488.

[17] M. Franz, J.S. McCarley, S. Roukos, Ad hoc and multilingual information retrieval at IBM, in: TREC 1998, pp. 104–115.

[18] E. Gabrilovich, S. Markovitch, Wikipedia-based Semantic Interpretation for Natural Language Processing, Journal of Artificial Intelligence Research (JAIR) 34 (2009) 443–498.

[19] M. de Gemmis, P. Lops, C. Musto, F. Narducci, G. Semeraro, Semantics-aware Content-based Recommender Systems, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, 2nd Edition, Springer, 2015, pp. 119–159.

[20] J. Giles, Internet Encyclopaedias Go Head to Head, Nature 438 (2005) 900–901.
[21] T. Gollins, M. Sanderson, Improving Cross Language Retrieval with Triangulated Translation, in: Proc. of the 24th Int. ACM SIGIR Conf. on Research and development in information retrieval 2001, ACM, pp. 90–95.
[22] F. Grosjean, Bilingual: Life and Reality, Harvard University Press, 2010.
[23] Z. Harris, Mathematical Structures of Language, New York: Interscience, 1968.
[24] J. Hu, L. Fang, Y. Cao, H. Zeng, H. Li, Q. Yang, Z. Chen, Enhancing Text Clustering by Leveraging Wikipedia Semantics, in: S. Myaeng, D.W. Oard, F. Sebastiani, T. Chua, M. Leong (Eds.), Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, ACM, 2008, pp. 179–186.
[25] P. Kanerva, Sparse Distributed Memory, MIT Press, 1988.
[26] M. Lesk, Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, in: Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC '86, ACM, 1986, pp. 24–26.
[27] M.L. Littman, S.T. Dumais, T.K. Landauer, Automatic Cross-language Information Retrieval using Latent Semantic Indexing, in: Cross-language information retrieval, Springer, 1998, pp. 51–62.
[28] P. Lops, C. Musto, F. Narducci, M. De Gemmis, P. Basile, G. Semeraro, Cross-language Personalization through a Semantic Content-based Recommender System, in: Artificial Intelligence: Methodology, Systems, and Applications, Springer, 2010, pp. 52–60.
[29] B. Magnini, C. Strapparava, Improving User Modelling with Content-based Techniques, in: Proc. 8th Int. Conf. User Modeling, Springer, 2001, pp. 74–83.
[30] C. Martinez-Cruz, C. Porcel, J. Bernabé-Moreno, E. Herrera-Viedma, A model to represent users trust in recommender systems using ontologies and fuzzy linguistic modeling, Information Sciences 311 (2015) 102 – 118.
[31] J.S. McCarley, Should We Translate the Documents or the Queries in Cross-language Information Retrieval?, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL '99, pp. 208–214.
[32] A. Moro, A. Raganato, R. Navigli, Entity Linking meets Word Sense Disambiguation: a Unified Approach, Transactions of the Association for Computational Linguistics 2 (2014) 231–244.
[33] C. Musto, F. Narducci, P. Basile, P. Lops, M. de Gemmis, G. Semeraro, Cross-Language Information Filtering: Word Sense Disambiguation vs. Distributional Models, in: Proceedigs of AI*IA 2011: Artificial Intelligence Around Man and Beyond - XIIth International Conference of the Italian Association for Artificial Intelligence, volume 6934 of *Lecture Notes in Computer Science*, Springer, 2011, pp. 250–261.
[34] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J.H.M. Korst, V. Pronk, R. Clout, Enhanced Semantic TV-Show Representation for Personalized Electronic Program Guides, in: J. Masthoff, B. Mobasher, M.C. Desmarais, R. Nkambou (Eds.), Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization, UMAP 2012, volume 7379 of *Lecture Notes in Computer Science*, Springer, 2012, pp. 188–199.
[35] J. Nadeau, C. Lointier, R. Morin, M. Descôteaux, Information Highways and the Francophone World: Current Situation and Strategies for the Future, in: INET'98 Conference: The Internet Summit, pp. 21–4.
[36] F. Narducci, M. Palmonari, G. Semeraro, Cross-Language Semantic Retrieval and Linking of E-Gov Services, in: H. Alani, L. Kagal, A. Fokoue, P.T. Groth, C. Biemann, J.X. Parreira, L. Aroyo, N.F. Noy, C. Welty, K. Janowicz (Eds.), The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Proceedings, Part II, volume 8219 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 130–145.
[37] R. Navigli, D. Jurgens, D. Vannella, Semeval-2013 task 12: Multilingual word sense disambiguation, in: Proc. of SemEval-2013, pp. 222–231.
[38] R. Navigli, S.P. Ponzetto, BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network, Artificial Intelligence 193 (2012) 217–250.
[39] J.Y. Nie, Cross-language Information Retrieval, Synthesis Lectures on Human Language Technologies 3 (2010) 1–125.
[40] D.W. Oard, P.G. Hackett, Document Translation for Cross-language Text Retrieval at the University of Maryland, in: Information Technology: The Sixth Text REtrieval Conference (TREC-6) 1997, pp. 687–696.
[41] F.J. Och, Statistical Machine Translation: Foundations and Recent Advances, in: Tutorial at Tenth Machine Translation Summit.
[42] C. Peters, M. Braschler, P. Clough, Multilingual Information Retrieval: from Research to Practice,

Springer, 2012.
[43] X.H. Pham, J.J. Jung, N.T. Nguyen, P. Kim, Ontology-based multilingual search in recommendation systems, Acta Polytechnica Hungarica 13 (2016).
[44] R. Chau and Chung-Hsing Yeh, Fuzzy Multilingual Information Filtering, in: FUZZ '03, 12th IEEE International Conference on Fuzzy Systems, pp. 767–771.
[45] D. Rao, P. McNamee, M. Dredze, Entity linking: Finding extracted entities in a knowledge base, in: T. Poibeau, H. Saggion, J. Piskorski, R. Yangarber (Eds.), Multi-source, Multilingual Information Extraction and Summarization, Theory and Applications of Natural Language Processing, Springer, 2013, pp. 93–115.
[46] J. Savoy, L. Dolamic, How effective is Google's Translation Service in Search?, Communications of the ACM 52 (2009) 139–143.
[47] P. Sorg, P. Cimiano, Exploiting Wikipedia for Cross-lingual and Multilingual Information Retrieval, Data & Knowledge Eng. 74 (2012) 26–45.
[48] B. Steichen, M.R. Ghorab, A. O'Connor, S. Lawless, V. Wade, Towards personalized multilingual information access - exploring the browsing and search behavior of multilingual users, in: V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, G. Houben (Eds.), Proceedings of the 22nd International Conference on User Modeling, Adaptation, and Personalization, UMAP 2014, volume 8538 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 435–446.
[49] V. Vehovar, Prospects of Small Countries in the Age of the Internet, Cyberimperialism? Global relations in the new electronic frontier (2001) 123–138.
[50] M.S. Wu, Modeling query-document dependencies with topic language models for information retrieval, Information Sciences 312 (2015) 1 – 12.