Cross-Lingual Recommendations in a Resource-Based Learning Scenario

Sebastian Schmidt, Philipp Scholl, Christoph Rensing, and Ralf Steinmetz

Multimedia Communications Lab - Technische Universität Darmstadt
64283 Darmstadt - Germany
Web: http://www.kom.tu-darmstadt.de
{sschmidt,scholl,rensing,ralf.steinmetz}@kom.tu-darmstadt.de

Abstract. CROKODIL is a platform supporting resource–based learning scenarios for self–directed, on–task learning with web resources. As CROKODIL enables the forming of possibly large learning communities, the stored data is growing in a large scale. Thus, an appropriate recommendation of tags and learning resources becomes increasingly important for supporting learners. We propose *semantic relatedness* between tags and resources as a basis of recommendation and identify Explicit Semantic Analysis (ESA) using Wikipedia as reference corpus as a viable option. However, data from CROKODIL shows that tags and resources are often composed in different languages. Thus, a monolingual approach to provide recommendations is not applicable in CROKODIL. Thus, we examine strategies for providing mappings between different languages, extending ESA to provide cross–lingual capabilities. Specifically, we present mapping strategies that utilize additional semantic information contained in Wikipedia. Based on CROKODIL's application scenario, we present an evaluation design and show results of cross–lingual ESA.

1 Introduction

With ever changing working environments and a decreased life-span of knowledge, learning becomes a lifelong process. The learning process which complements institutional education (including school, apprenticeship, university etc.) is characterized by the learner's self-responsibility and self-monitoring. Learning materials are available and accessible (e.g. on the Web), but not necessarily prepared for learning by a teacher like in traditional learning environments. Self-directed learning using learning materials is called *Resource–Based Learning* (RBL). In *RBL* settings, a major challenge for learners is finding relevant *Learning Resources* (LRs). One common strategy for this form of learning is to use a web search engine or specialized digital libraries. In learning settings, however, where a community like a learning group, a class of students or a group of colleagues does already exist, the probability that other members of the community have already found relevant information is high. In order to discover this information, recommender systems that recommend information based on different features can be useful.

In this paper, we examine the data from a user study with the *RBL* platform *CROKODIL* with special regard to the language of the stored information (Section 2), analyze existing approaches for semantic relatedness with regard to cross-linguality (Section 3) and propose a feasible approach to enable cross-lingual recommendations (Section 4). Further, we evaluate this approach within our scenario (Section 5) and give a conclusion and prospect on further work (Section 6).

2 Multilinguality in the Usage Scenario

2.1 CROKODIL

The e–learning platform *CROKODIL* supports learners in finding, collecting and organizing learning materials from the Web in *RBL*. All data within the platform that is inserted by the users is stored in information items. Those can be *LR*s, learners, tags or various other types. Information items can be interconnected via relations. The overall resulting graph represents a kind of folksonomy. An *LR* in *CROKODIL* can be a whole document (web-page, pdf-file, textfile, etc.) or a short text fragment. Learners have a profile describing the represented person. The tags that are used by learners to describe the *LR*s are terms consisting of one or several words.

CROKODIL uses a recommendation engine that attempts to provide tags or *LR*s that are likely to be of interest to the learner. These recommendations bridge the gap between the information need of the learners and already existing possibly matching information items in the system. However, up to now *CROKODIL* only provides recommendations based on structural properties of the network built by information items and relations, e.g. whether there are explicit connections between two *LR*s over a defined set of tags. This means that if there is no explicit relation between two *LR*s, the recommendation engine is not able to infer this connection. Therefore, the formation of separate partitions of the networks is favoured, especially as different users commonly have a different terminology for denoting related information (eg. "Technology Enhanced Learning" and "e–learning"). Information items containing identical or almost identical semantic information to those the user has already stored are only of minor interest. Therefore, we aim to enable recommendation of semantically *related* information items.

Another challenge that *CROKODIL* has to meet with regard to recommending relevant items is that the overall knowledge base is expected to be sparse. In contrast to social bookmarking applications like Delicious, *CROKODIL* does not have millions of users and therefore collaborative filtering [1] might not be appropriate for recommending items. Therefore, a content–based recommendation paradigm is targeted in our work.

2.2 Language of Tags and Learning Resources

We examined the data from a user study [2] with ELWMS.KOM, a predecessor of *CROKODIL*, in order to determine the used language of *LR*s and tags.

In the study, 21 knowledge workers at Technische Universität Darmstadt used ELWMS.KOM over a period of several weeks. Table 1 shows the language distribution of the stored web resources. A majority of the resources are in English, which is contradictory to the local language and the mother tongue of most of the participants (17 of the participants are German native speakers).

Table 1. Web resources contained in the knowledge base grouped by language and fraction of resource language chosen by German and non–German native speakers.

Language	Web Resource Count	Web Resource Percentage	by German native speakers (4)	by Non–German native speakers (17)
English	333	75.33%	73.31%	79.45%
German	98	22.18%	23.99%	18.49%
French	2	0.46%	-	1.37%
Page forbidden (403)	1	0.22%	-	0.69%
Page unavailable (404)	8	1.81%	2.70%	-
Total	442	100.00%	100.00%	100.00%

Further, the language of the tags was evaluated. The results (cf. Table 2) let us infer that the language of a resource does not necessarily correspond to the language of the attached tags (e.g. French is not contained in the tag languages). Finally, in 70.2% of the cases the languages of tags and tagged resources correspond.

Table 2. Tags used for web resources in different languages in *ELWMS.KOM* sample. Note that German and non–German native speakers were involved and the number of participants does not match the numbers in the resource language experiment, as only 18 participants applied tags to resources.

Type	Tag Count	Tag Count in %	by German native speakers (15)	by Non–German native speakers (3)
English	300	30.70%	25.40%	42.91%
German	183	18.73%	22.17%	10.81%
Ambiguous Language	194	19.87%	20.41%	18.58%
Named entity (uni–lingual)	240	24.56%	28.63%	15.20%
Date or year	60	6.14%	3.39%	12.50%
Total	977	100.00%	100.00%	100.00%

This analysis shows that in real–world settings, the usage of *RBL* often crosses language borders. For content–based recommendations, this adds a dimension of complexity, as the language of *LR*s and tags has to be taken into account additionally. The learners' choice of tags is often influenced by the language a *LR* is composed in. This means that the same learner could use different tags for the same concept, eg. one user tagged related *LR*s with the English "visual" and the German "Visualisierung". This adds to the aforementioned challenges.

3 Related Work

Content-based recommendations can be performed by analyzing the semantic content of the available information items and by recommending those with the highest semantic closeness to other items which are known to be relevant for the user. According to Budanitsky and Hirst [3], there is a considerable difference between the two notions of semantic closeness *semantic similarity* and *semantic relatedness*. Semantic similarity denotes the degree of two different terms describing the same concept, e.g. the terms "cash" and "dough" have a high semantic similarity, because "dough" is a colloquial synonym for "cash". In contrast, semantic relatedness mimics the associative perception of humans. E.g. the terms "cash" and "bank" do not have a semantic similarity, but are semantically related because they often occur in a common context. Especially in our scenario, the concept of semantic relatedness is appropriate as recommendation of information items with a related content is targeted.

3.1 Semantic Relatedness

Milne and Witten [4] state that "any attempt to compute semantic relatedness automatically must also consult external sources of knowledge". Thus, all approaches to determine semantic relatedness utilize additional information by employing reference corpora in order to provide additional general knowledge. WordNet [5] is often used as an external source of knowledge to enable the calculation of semantic relatedness [3, 6]. However, in recent work the focus has shifted to Wikipedia as a knowledge base because of its corpus size and its up-to-dateness.

Explicit Semantic Analysis (ESA) A promising approach for calculating semantic relatedness called *ESA* has been proposed by Gabrilovich and Markovitch [7]. Here, documents are not represented by means of terms but by their similarity to concepts derived from a reference collection of documents. *ESA* is based on the assumption that in the reference document collection, an article corresponds to a semantically distinct concept. Thus, by comparing documents based on their terminology to all articles in the document collection that have been pre–processed by tokenization, stemming, stop word removal and a term weight metric, a vector is obtained that contains a similarity value to each of the articles. This process step is called *semantic analysis*. The vector, called *semantic interpretation vector*, abstracts from the actual term occurrences and thus represents a semantic dimension of that document. A major advantage of *ESA* is that semantic relatedness can be calculated for terms and documents alike, providing good and stable results for both modes [7].

Formally, the document collection is represented as a matrix M with the dimensions $n \times m$ (called *semantic interpreter*), where n is the number of articles and m the number of occurring terms in the corpus. M contains (normalized) *TF–IDF* document vectors of the articles.

For calculating the similarity between the document and the corpus, the *cosine similarity measure* is employed. Analogously, two documents represented as semantic interpretation vectors can be easily compared by using cosine similarity again.

Despite its primary usage with Wikipedia, *ESA* is also applicable to different reference corpora. Gabrilovich and Markovitch have used the *Open Directory Project* (ODP) as well as Wikipedia, showing that *ESA* using Wikipedia outperforms the *ODP* reference corpus. They state that Wikipedia is especially practical for *ESA* as each of Wikipedia's articles ideally describes one concept.

Gabrilovich and Markovitch show that *ESA* using Wikipedia as a reference corpus outperforms other approaches like WikiRelate! [8], approaches based on WordNet [3] or Roget's Thesaurus [9] and *Latent Semantic Analysis* [7].

Extensions for ESA, which consider Wikipedia's internal link structure and its category system have been proposed and successfully evaluated [10]. Thus, further usage of structural information obtained from Wikipedia seems to be promising for the calculation of semantic relatedness.

3.2 Cross-Language Semantic Relatedness

As the Web contains documents composed in a variety of languages the need for cross-lingual approaches in semantic relatedness determination emerges. Semantic relatedness across language borders has therefore become a focus of research in recent years.

Schönhofen et al. [11] investigate the usage of Wikipedia for *Cross-Language Information Retrieval* (CL IR), aiming to query and retrieve English documents by German and Hungarian queries. For that purpose they first do a "word-by-word translation by dictionary", yielding in many cases a large set of word pairs for a single word in the source and the possible translations in the target language. In order to overcome this issue, they first aim to maximize the bigram similarity between the different translation combinations of adjacent words, consulting statistics obtained from the English Wikipedia as a reference corpus in the target language. Then, the links between pairs of articles containing the two translated terms in the article title are used to rank the translations. After having obtained the ranks for the translation pairs, Schönhofen et al. combine both measures to a final rank which results in an order describing the most probable terms. Although this approach benefits from the networked structure of Wikipedia which mirrors the semantic relatedness of concepts, it is still a term based approach which does not take the global term distribution, a measure of global term relevance, into account.

Potthast et al. [12] and Sorg et al. [13] both investigate a usage of *ESA* in cross-lingual contexts. Potthast et al. focus on the field of automatic cross-lingual plagiarism detection. They consider a language-independent concept space to which for each supported language a document-collection is aligned via a one-to-one mapping. This requires a reference corpus which contains articles describing the same set of concepts in different languages. Hence, only a subset of

articles can be considered for the semantic relatedness computation in the case of Wikipedia usage. Because of their assumption of a bijective article mapping function and their restrictive usage of disjunction of all articles, the direction of their mapping does not matter to the results. Sorg et al. present a slightly more elaborated approach which does not assume a one-to-one mapping between articles in the corpus but a many-to-one mapping for articles in the source language to articles in the target language. So each target article might be targeted from different articles in the source language. In their approach, they first compute the ESA vector in the source language and map it to the target language afterwards by summing up the relatedness values from all concepts in the source language pointing to a single concept in the target language. They indicate a a good correlation to human rankings. But in case of topics being underrepresented in a specific Wikipedia language version, their approach is not capable of mapping the interpretation vectors without loosing semantic information.

Dumais et al. [14] present an approach using *Latent Semantic Indexing* (LSA) for *CL-IR* with a common space containing terms from different languages and show to perform well in CL-IR tasks. In contrast to *ESA*, *LSI* does not allow to give human-readable explanations about the reasoning for a particular relatedness rating.

4 Our Approach

Due to the promising results of ESA in mono- and multilingual settings we decided to apply this method in *CROKODIL* and extend it by an approach which can handle underrepresented topics in the Wikipedia. This challenge has not been addressed within the work presented in the previous section. Therefore, in the following, adaptations made to *ESA* in order to use it in multilingual environments are explained. Steps 1, 2, 3 and 5 are basically the steps executed in the original ESA procedure. The index l indicates the respective language.

1. The reference corpus (in our case the Wikipedia in language l) is preprocessed with stemming, stopword removal, article filtering[1], *TF–IDF* calculation and normalization. The result is a language-specific matrix M_l, the *semantic interpreter*, with the shape $n_l \times m_l$, where n_l is the number of articles and m_l the number of terms.
2. For each document d_l (in language l) that is to be compared, the same preprocessing steps have to be executed, so that the result is the document vector v_d^l with the form $m_l \times 1$, where m_l is the number of terms.
3. As all document vectors are normalized, the interpretation vector $i_{esa}^l = M_l \cdot v_d^l$ that represents the cosine similarity of v_d^l with all article vectors of M_l is simply computed by applying the inner product with the matrix M_l. The result is the interpretation vector i_{esa}^l with the dimensions $1 \times n_l$.

[1] In previous work [15] the systematic filtering of articles has shown to lessen the concept space without impairing *ESA*'s quality.

4. All interpretation vectors $i_{esa}^{l_s}$ now have to be transferred (mapped) from their respective source language space into one common target language space l_t and result in vectors $i_{esa}^{l_t}$. We will explain two different mapping strategies later on.
5. Finally, the resulting interpretation vectors $i_{esa}^{l_t}$ can be compared pairwise using the cosine similarity.

The first step has to be executed for all languages which are supported (in our case German and English). Step 4 is not contained in the original ESA implementation. It needs to be executed in our cross-lingual approach because step 5 requires the interpretation vectors to have the same dimensions. A mapping has to be applied to equalize the number of dimensions of the interpretation vector (corresponding to the number of Wikipedia articles in the respective language). This mapping is the crucial step in cross-lingual semantic relatedness calculation with ESA. The more concepts can be mapped from one language to the other, the closer the quality of interlingual semantic relatedness matches that of monolingual ESA. We now describe the two mapping techniques we apply.

The recommendation quality is expected to be constant over the time. Preliminary experiments have shown that the use of updated Wikipedia versions does not have a major influence on the computation's quality, only in case of topics not being represented in the Wikipedia up to now.

4.1 Direct CL Mapping

Wikipedia articles can be interlinked to corresponding articles in another language version of the Wikipedia by *interlanguage links* (called ILLs). Ideally, the interlinked articles describe exactly the same concept. If such a link exists for the article whose dimension needs to be mapped, it is used and the weight (determined in step 3 of the ESA algorithm) is transfered to the target article of the link. This matches the approach presented by Potthast et al. and Sorg et al. [12, 13].

If for each article in the source language a linked article in the target language exists, the mapping can be expected to be ideal. Unfortunately, this is usually not the case. Considering the German and the English Wikipedia (cf. Table 3) the total numbers of articles differ strongly, thus, a one-to-one article mapping can not be achieved. For example, only around 18% of the English articles have a (linked) German correspondent, thus 82% of the concept dimensions would be discarded during the mapping step using only *ILL*s.

Table 3. Key figures of the size of the English and the German Wikipedias

	English Wikipedia	German Wikipedia
Date of Dump	2011-01-15	2010-06-03
Number of articles	3,601,228	1,095,678
*CLL*s to other language	657,874	606,160

4.2 Meta CL Mapping

We present a novel approach which aims at overcoming CL mapping issues due to articles not being interlinked to other languages. First, all weights in $i^{l_s}_{esa}$ of articles a_{l_s} being interlinked via a *ILL* to an article a_{l_t} are directly transferred into the new vector $i^{l_t}_{esa}$. Afterwards, for each article a_{l_s} which does not have an outgoing *ILL*, the set of articles within the same language version of the Wikipedia that are interlinked from a_{l_s} are considered. For all those articles, the articles which have an *ILL* to an article in the target language are taken into account. The resulting set of articles in the target language represents a *meta concept* for the article a_{l_s}. The link between a_{l_s} and the meta concept is called in the following *meta interlanguage link* (MILL). The weight for a_{l_s} is equally distributed to all elements in the *meta concept*. Figure 1 shows exemplarily the *MILL* for the English article "Clark Kent" to its meta concept in the German Wikipedia. The original weight 0.8 is divided and each of the four elements in the meta concept get assigned a weight of 0.2.

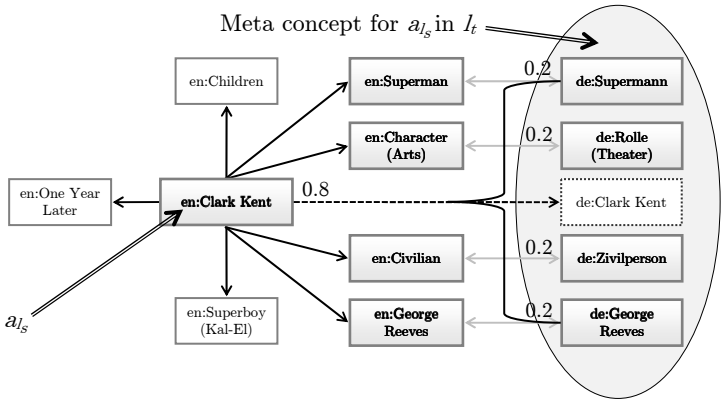


Fig. 1. Example of a *meta interlanguage link* from the English article *Clark Kent* that does not have a corresponding article in German

5 Evaluation

5.1 Corpora

In order to evaluate our approach we needed an appropriate evaluation corpus. Following the considered usage scenario for our approach, we identified the following requirements:

- The main languages used within the *CROKODIL* system in the user study were English and German. Thus, the evaluation corpora should consist of documents in those languages.

- The information items to recommend are tags and multi-term documents. Thus, we need corpora containing terms and multi-term documents.
- In most cases, nouns have been used as tags, thus the corpus containing terms should include primarily nouns.

We observed, that often *LR* and their assigned tags share the same language (cf. Section 2.2). Thus, cross-recommendation between tags and *LR*s are not focus of our work because this is mainly a monolingual task. The term-term relatedness and the document-document relatedness are evaluated separately due to the different conditions and strongly differing results for those two evaluations in other studies [10]. For the different scenarios, we use different corpora which are explained in the following.

The Document Corpus To the best of our knowledge, there exists no appropriate multilingual document corpus that contains relatedness values for the documents. We decided to focus on a parallel corpus, containing exactly similar documents in different languages. The corpus *Europarl test2007*[2] is a testing subset of the *Europarl* [16] corpus containing 2000 parallel sentences in four languages. The content of the *Europarl* corpus has been extracted from proceedings of the European Parliament. For our evaluations, we use a subset of the English-German parallel corpus with 300 documents.

The Term Corpus The multilingual dataset *Schm280* [15] is adapted from the English *WordSim353* dataset created by Finkelstein et al. [17]. It contains 280 English word pairs (mainly nouns) with their German equivalents translated by 12 participants, each value pair with a value of semantic relatedness (between 0.0 and 10.0) rated by at least 13 subjects. The words are very generic and not restricted to a single topic.

5.2 Evaluation Methodology and Results

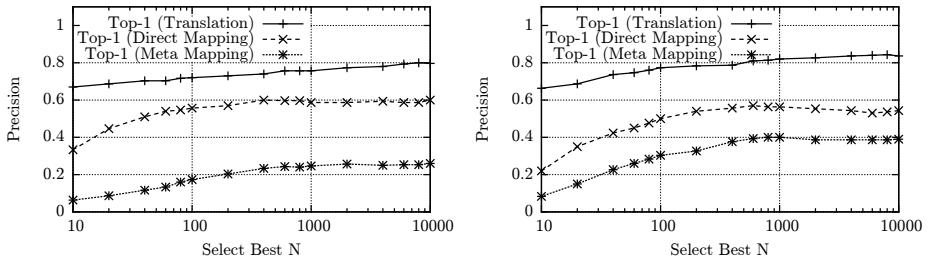
In order to evaluate our approach we execute different experiments. We aim to reduce storage space for the created interpretation vectors and to minimize the computation time for the cosine similarity computation. Therefore, we define and apply the `selectBestN` function [10, 13] which is aimed at choosing only the most relevant concept dimensions (in dependency of their weights). N should be chosen as small as possible without having a considerable negative impact on the results. Therefore, all evaluations are executed with a varying number of dimensions of the interpretation vectors. We parametrize the function with 16 values for N between 10 and 10000.

Document-Document Similarity Using an Information Retrieval scenario we evaluate the applicability on cross-lingual document similarity. Our system is

[2] available via http://www.statmt.org/wmt07/shared-task.html, retrieved 2011-04-04

queried with a sentence from the *Europarl test2007* corpus in one language and the parallel document from the other language is expected as result. Because one document has exactly one correspondent document in the other language, a *Top–k* evaluation is applicable in this scenario. The scenario is evaluated using both the *ILL* and *MILL* mapping to map the interpretation vectors of the documents to the query language l_t. Additionally for comparison, an evaluation with monolingual ESA is applied where all documents are in advance translated to the respective query language using Google Translator[3].

In the evaluation the mapping approaches do not seem to achieve a similar precision like the comparison approach using monolingual ESA with Google Translator (cf. Figure 2). The correct English document is returned as the best ranked one when passing a German query in at most 80% of the cases for the translational approach, in 60% for the *ILL* mapping approach and in 26% for the *MILL* approach. The correct German document is returned as the best ranked one when passing an English query in at most 84% of the cases for the translational approach, in 57% for the *ILL* approach and in 40% for the *MILL* approach. Table 4 shows a detailed analysis for different parameterizations of k. Shown is



(a) German target language space (b) English target language space

Fig. 2. Precision for Top-1 evaluation

the maximal precision for each approach over all parameterizations of N. According to the previous analysis for $k = 1$, the monolingual setting is always superior to the *ILL* and the *MILL* mapping. Further, the *MILL* mapping can not achieve a precision similar to the results of the *ILL* mapping.

Term-Term Relatedness In order to determine the statistical dependence of pairwise relatedness results, we use the Spearman's rank correlation, also known as Spearman's rho [18]. It considers ranks of the single values instead of the values itself as the Pearson correlation does. As for relatedness measures, the ranking of values is more important than the exact relatedness value, we

[3] http://translate.google.de/, retrieved 2011-04-04

Table 4. Maximum precisions for different k in *Top-k* evaluation and considered number of dimensions of the interpretation vectors

		$k=1$		$k=5$		$k=10$	
German to English	Translation	0.84	8000	0.94	2000	0.96	1000
	ILL mapping	0.57	600	0.82	1000	0.88	4000
	MILL mapping	0.4	800	0.72	4000	0.83	2000
English to German	Translation	0.8	8000	0.92	8000	0.92	4000
	ILL mapping	0.6	10000	0.8	2000	0.87	800
	MILL mapping	0.26	10000	0.54	6000	0.64	2000

prefer this measurement. Other authors also used this procedure to evaluate their semantic relatedness approaches (i.e. [12]). In our evaluation, we use this approach to determine how close to human relatedness rankings the rankings determined by our automatic relatedness determination techniques are.

The significance of the difference between two correlations can be determined by using t_{diff} [18]. It is used to check whether the correlation between the pairs of variables (x, y) and (z, y) is significantly different. It is defined as:

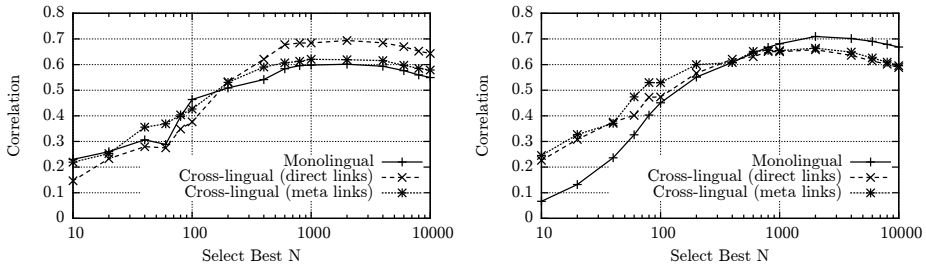
$$t_{\text{diff}} = (\rho_{xy} - \rho_{zy}) \sqrt{\frac{(k-3)(1+\rho_{xz})}{2(1-\rho_{xy}^2 - \rho_{xz}^2 - \rho_{zy}^2 + 2\rho_{xy}\rho_{xz}\rho_{zy})}} \tag{1}$$

The resulting values for t_{diff} are compared with the critical values of the t–distribution.

The evaluation is executed using the *Schm280* corpus to determine the rank correlation between the relatedness values determined by humans and our approach. Within the evaluation of the cross-lingual experiments, the relatedness is determined for each pair with the first word being in language l_t and the second word being in language l_s and mapped to l_t. As a comparison, we determine the monolingual correlation with both words existing in l_t's language space.

Figure 3 shows Spearman's rank correlation ρ between relatedness determined by human raters and relatedness determined by the *ILL* and *MILL* mappings. Further, results of the monolingual setting are shown.

In the monolingual setting, the Spearman's rank correlation coefficient ρ is maximal for *selectBestN* with $N \approx 2000$ for both languages (with $\rho_{en}(i_{esa}^{2000}) = 0.71$ and $\rho_{de}(i_{esa}^{2000}) = 0.60$). In cross–lingual settings, the experiment in the English language space l_{en} outperforms the experiment using the German language space by approximately 18% in terms of correlation. When mapping the second term of each term pair from German to English by *ILL*, the correlation is significantly lower than the monolingual experiment (for $\rho_{en}(i_{esa}^{2000}) = 0.66$, $t_{\text{diff}} = 2.01$, $p < .05$). In the German language space however, the *ILL* mapping results in a significantly higher correlation (with $\rho_{de}(i_{esa}^{2000}) = 0.69$, $t_{\text{diff}} = 3.37$, $p < .01$) compared to the monolingual approach. This shows that the quality of *ESA* is highly dependent on the used language space and a cross–lingual mapping transfers qualitative properties to the target language space. Remarkable is the slight improvement for *MILL* mapping compared to *ILL* mapping and monolingual evaluation for small N. We explain the better performance of *MILL* compared



(a) German target language space (b) English target language space

Fig. 3. Correlation for Semantic Relatedness between human judgement and our approach

to *ILL* mapping with an accumulation of article links to relevant concepts (which usually are general and therefore have a statistically higher probability of being the target of article links), boosting the values for these relevant concepts due to spread. In fact, an analysis of the data shows that the relevant concepts benefit from the incomming *MILL*s. This effect can be used when N has to be set to a low value for performance and storage reasons. In the English language space, *ILL* and *MILL* mappings perform similarly for large N. However, especially in the German language space, the correlation of *MILL* mapping is lower for high values of N.

6 Conclusions and Further Work

In this paper, we presented a multilingual scenario of resource-based learning using web resources. *CROKODIL* enables users to store various types of learning resources and assign tags to them without being confined to a single language. In order to provided content-based recommendations across language borders we proposed an adaption to Explicit Semantic Analysis which does not only take interlanguage links into account but additionally Wikipedia's internal link structure: the *MILL* mapping. In the evaluation of term-term relatedness, it showed better correlation for low dimensional interpretation vectors, even compared to the monolingual evaluation. Thus, due to the reduction of computational complexity the *MILL* mapping can be helpful for real-time applications like the on-line recommendation of newly integrated information items in *CROKODIL*. Further evaluation did not show improvements when compared to the mapping using only *ILL*s. We however still see potential of this approach, especially regarding underrepresented topics in the Wikipedia. When topics are underrepresented for a certain language, documents covering those can not be mapped into this language space using only *ILL*s. This can lead to a loss of important information if strongly weighted concepts are discarded. For document-document

similarity tasks we currently recommend the usage of *ILL* mapping due to the restrictive usage policy of the Google API which does not allow the translation-based approach. The goal of document recommendations in *CROKODIL* is not to recommend similar resources but rather semantically related resources, which is not directly addressed by our evaluation. We imagine the following scenario when including cross-lingual content based recommendation in *CROKODIL*:

1. The users insert learning resources and attach tags to them, both can be in any of the supported languages.
2. For each inserted learning resource and tag, a semantic interpretation vector is created, mapped to a common language, reduced via `SelectBestN` and stored in the system.
3. Users get those tags/learning resources recommended which have the highest semantic relatedness to the tags/learning resources they have used.

In future work we will focus on further adaptions and evaluations of our approach as we consider the achieved results as motivation for further research in this field. By distributing the weight of a mapped concept well-directed to the different elements of the *meta concept*, we hope to enable the amplification of important linked concepts. Further, if only strongly weighted concepts in the original interpretation vectors are mapped via *MILL*, noise might be attenuated. In order to satisfy the requirements for the described application area more precisely, an evaluation with further corpora is needed. In addition for the evaluation of the recommendation of documents, a corpus containing relatedness values is required. As we have seen [10], approaches relying on Wikipedia's link structure can outperform plain *ESA* in scenarios where documents are rather related than they are similar. Considering the recommendation of tags, a more specific corpus would represent the scenario better. The used corpus contains highly generic terms, which are strongly represented in Wikipedia and thus the approach relying on meta concepts cannot demonstrate its benefits. Finally, for adapting the approach to general cross-lingual recommendation, evaluations with further languages are needed. Especially in scenarios with languages of considerably different structure (like English and Arabic) interesting results can be expected as our mapping approach only relies on the quality of the Wikipedia reference corpus in the respective language and not the quality of a translation engine.

Acknowledgments. This work was supported by funds from the German Federal Ministry of Education and Research under the mark 01 PF 08015 A and from the European Social Fund of the European Union (ESF). The responsibility for the contents of this publication lies with the authors.

References

1. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems: An Introduction. First edn. Cambridge University Press (2010)

2. Böhnstedt, D., Scholl, P., Rensing, C., Steinmetz, R.: Enhancing an Environment for Knowledge Acquisition based on Web Resources by Automatic Tag Type Identification. In: Proceedings of International Conference on Computer-aided Learning 2010 (ICL 2010), Kassel, Kassel University Press (Sep 2010) 380–389
3. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics **32**(1) (2006) 13–47
4. Milne, D., Witten, I.H.: An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In: Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA (2008) 25–30
5. Fellbaum, C.: Wordnet: An Electronic Lexical Database. MIT Press (1998)
6. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference Research on Computational Linguistics (ROCLING X). (1997)
7. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. (2007) 6–12
8. Strube, M., Ponzetto, S.P.: WikiRelate! Computing Semantic Relatedness Using Wikipedia. In: Proceedings of the National Conference on Artificial Intelligence. Volume 21., Menlo Park, CA; Cambridge, MA; London;, AAAI Press; MIT Press (2006) 1419ff
9. Jarmasz, M., Szpakowicz, S.: Roget's Thesaurus and Semantic Similarity. Recent Advances in Natural Language Processing III: Selected Papers from RANLP **1** (2004) 111
10. Scholl, P., Böhnstedt, D., Domínguez García, R., Rensing, C., Steinmetz, R.: Extended Explicit Semantic Analysis for Calculating Semantic Relatedness of Web Resources. Proceedings of EC-TEL: Sustaining TEL: From Innovation to Learning and Practice (2010) 324–339
11. Schönhofen, P., Benczúr, A., Bíró, I., Csalogány, K.: Cross–Language Retrieval with Wikipedia. Advances in Multilingual and Multimodal Information Retrieval **5152** (2008) 72–79
12. Potthast, M., Stein, B., Anderka, M.: A Wikipedia–Based Multilingual Retrieval Model. In: Proceedings of the IR research, 30th European conference on Advances in Information Retrieval, Springer-Verlag (2008) 522–530
13. Sorg, P., Cimiano, P.: Cross–lingual Information Retrieval with Explicit Semantic Analysis. In: Working Notes for the CLEF 2008 Workshop. (2008)
14. Dumais, S., Letsche, T., Littman, M., Landauer, T.: Automatic Cross-Language Retrieval Using Latent Semantic Indexing. In: AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. (1997) 15–21
15. Schmidt, S.: Language-Independent Semantic Relatedness Using Wikipedia as Reference Corpus. Master's thesis, Technische Universität Darmstadt (2010)
16. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Machine Translation Summit 2005. Volume 5., European Association for Machine Translation (Sep 2005) 79–86
17. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing Search in Context: The Concept Revisited. ACM Transactions on Information Systems (TOIS) **20**(1) (Jan 2002) 116–131
18. Field, A.P.: Discovering Statistics Using SPSS. SAGE publications (2009)