Projective methods for mining missing translations in DBpedia

Laurent Jakubina
RALI - DIRO
Université de Montréal
jakubinl@iro.umontreal.ca

Philippe Langlais
RALI - DIRO
Université de Montréal
felipe@iro.umontreal.ca

Abstract

Each entry (concept) in DBpedia comes along a set of surface strings (property rdfs:label) which are possible realizations of the concept being described. Currently, only a fifth of the English DBpedia entries have a surface string in French, which severely limits the deployment of Semantic Web Annotation for this language. In this paper, we investigate the task of identifying missing translations, contrasting two projective approaches. We show that the problem is actually challenging, and that a carefully engineered baseline is not easy to outperform.

1 Introduction

The LOD (Linked Open Data) (Bizer et al., 2009) is conceived as a language independent resource in the sense that the information is represented by abstract concepts to which "human-readable" strings — possibly in different languages — are attached, *e.g.* the rdfs:label property in DBpedia. For instance, we can access the abstract concept of computer by natural language queries such as ordinateur (rdfs:label@fr) in French or computer (rdfs:label@en) in English. Thanks to this, Semantic Web offers the advantage of having a truly multilingual World Wide Web (Gracia et al., 2012).

At the core of LOD, lies DBpedia (Jens Lehmann, 2014), the largest dataset that constitutes a hub to which most other LOD datasets are linked. [1] Since DBpedia is (automatically) generated from Wikipedia, which is multilingual, one would expect that each concept in DBpedia is labeled with a French surface string. This is for instance the case of the concept House of Commons of Canada[2] which is labeled in French as Chambre des communes du Canada. One problem, however, is that most labels are currently in English (Gómez-Pérez et al., 2013).

Indeed, the majority of datasets in LOD are primarily generated from the extraction of anglophone resources. DBpedia, the endogenous RDF dataset of Wikipedia is no exception here, since it proposes labels in French (rdfs:label@fr) for only one fifth[3] of the concepts. Of course, all concepts in English Wikipedia have at least one English label. For instance, the concept School life expectancy[4] has — at least at the time of writing — no label in French, while for instance, durée moyenne de scolarité appears in the (French) article Indice_de_développement_humain,[5] and is a good translation of the English term.

This situation comes from the fact that currently, a concept in DBpedia receives as its rdfs:label property in a given language the title of the Wikipedia article which is inter-language linked to the (English) Wikipedia article associated to the DBpedia concept.

The lack of surface strings in a foreign language does not only reduce the usefulness of RDF indexing engines such as sig.ma,[6] but also limits the deployment of Semantic Web Annotator (SWA) systems; *e.g.* (Mihalcea and Csomai, 2007; Milne and Witten, 2008). This motivates the present study, which aims at automatically mining French labels for the concepts in DBpedia that do not

[1] December 2014 - http://lod-cloud.net/

[2] http://dbpedia.org/page/House_of_Commons_of_Canada

[3] http://wiki.dbpedia.org/Datasets/DatasetStatistics

[4] http://dbpedia.org/page/School_life_expectancy

[5] http://fr.wikipedia.org/wiki/Indice_de_developpement_humain

[6] http://sig.ma

possess one yet.

Identifying the translations of (English) Wikipedia article titles is partially solved in the BabelNet project (Navigli and Ponzetto, 2012). In this project, the translation of concepts in Wikipedia that are not inter-language linked are taken care of by applying machine translation on (minimum 3 and maximum 10) sentences extracted from Wikipedia that contain a link to the article whose title they seek to translate. The most frequent translation is finally selected. There are on the order of 500k articles in English Wikipedia that do not link to an article in French and which are not named entities (which typically do not require translation). BabelNet[7] provides a translation (not necessarily a good one) for 13% of them. This suggests that the projection of a resource such as `DBpedia` into French is not yet a solved problem.

In the remainder, we describe the approaches we tested in Section 2. Our experimental protocol is presented in Section 3. Section 4 reports the results we obtained. We conclude in Section 5.

2 Approaches

Identifying the translations of a term in a comparable corpus — two texts (one in each language of interest) that share similar topics without being in translation relation — is a challenge that has attracted many researchers. See (Sharoff et al., 2013) for a recent overview of the state-of-the-art in this field. In this work, we investigated several variants of two approaches for extracting translations from a comparable corpus: the seminal approach described in (Rapp, 1995) which uses a seed bilingual lexicon to induces new translations, and the approach of Bouamor et al. (2013) which instead exploits the Wikipedia structure. The latter approach has been shown to outperform the former significantly on a task of translating 110 terms in 4 different domains, making use of medium-sized corpora.[8]

2.1 Standard Approach (STAND)

The idea that the context of a term and the one of its translation share similarities that can be used to rank translation candidates has been previously investigated in (Rapp, 1995; Fung, 1998). Since then, many variants of this idea have been tested; see (Sharoff et al., 2013) for a recent discussion.

	w_1	$\neg w_1$	
w_2	O_{11}	O_{12}	R_1
$\neg w_2$	O_{21}	O_{22}	R_2
	C_1	C_2	N

Table 1: Contingency table

We reproduced this approach in this work. In a nutshell, each term to be translated is represented by a so-called *context vector*; that is, the set of words that co-occur with this term in the source part of the corpus. An *association measure* is typically used to score the strength of the correlation between the term and the context words. Each translation candidate (typically each word of the target vocabulary) is similarly represented in the target language. Thanks to a *bilingual seed lexicon*, the source context vector is projected into a target one.[9] This projected target language vector is then compared to the vector of each of the target language candidates by the means of a *similarity measure*.

There are several parameters to the approach among which the size of the window used to collect co-occurrent words, the association and the similarity measures, as well as the seed lexicon.

We investigate the impact of the window size in section 4. We also compare two different association measures, namely the discontinuous odds-ratio (Evert, 2005, p. 86) named ORD hereafter, and the log-likelihood ratio (Dunning, 1993), named LLR, the most popular measures used in this line of work. Both measures (Eq. 1 and 2) are computed directly from the (monolingual) contingency table depicted in Table 1 for two words w_1 and w_2 where, for instance, O_{12} stands for the number of times w_1 occurs in a window, while w_2 does not.

$$\text{ORD}(w_1, w_2) = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})} \quad (1)$$

$$\text{LLR}(w_1, w_2) = 2 \sum_{ij} O_{ij} \log \frac{N \times O_{ij}}{R_i \times C_j} \quad (2)$$

[7] Version 2.0.1 - March 2014

[8] 400k words on the English side, 260k words on the French side.

[9] In our implementation, when no translation is found for a source word, the word is left as such in the target context vector. On the contrary, multiple translations are all added to the target context vector.

We did not investigate the impact of the nature and size of the bilingual seed lexicon, but decided to use one large lexicons comprising 116 354 word pairs populated from several available resources as well as an in-house bilingual lexicon.[10] A similar choice is made in (Bouamor et al., 2013) where a seed lexicon of approximately 120 000 entries is being used, and in (Hazem et al., 2013), where the the authors use a lexicon of 200 000 entries (before preprocessing).

Since in (Laroche and Langlais, 2010) the best performing variant uses the cosine similarity measure (Eq. 3), we used it in our experiments.[11]

$$cos(v_{src}, v_{trg}) = \frac{v_{src} \cdot v_{trg}}{\|v_{src}\| \cdot \|v_{trg}\|} \quad (3)$$

In the standard approach, the co-occurrent words are extracted from all the source documents of the comparable corpus in which the term to translate appears. We name this variant STAND hereafter.

2.2 Neighbourhood variants (LKI, LKO, CMP and RA)

Since we are interested in translating Wikipedia titles, a natural way of populating the context vector of a term is to consider the occurrences of this term in the article whose title we seek to translate. This avoids populating the context vector with words co-occurring with different senses of the word to translate. We implemented such a variant which is inherently facing the issue that too few occurrences of the term of interest may appear in a single article, especially in our case where the average length of a Wikipedia article is approximatively 1 400 words. Therefore we considered a variant which involves a *neighbourhood function*, that is, a function that returns a set of Wikipedia articles related to the one under consideration for translation. We investigated three such functions (as well as many combinations of them):

LKI(a) returns the set of articles that have a link pointing to the article a under consideration (in links). For instance, both `Computer_Science` and `Art` are two articles pointing to `Entertainment`.

LKO(a) returns the set of articles to which a points to (out links). For instance the article `Entertainment` points to `Party` and `Fun`.

CMP(a) returns the set of articles that are the most similar to a. We used the *MoreLikeThis* method of the search engine Lucene[12] for this. For instance, `Dance` and `Dance in Indonesia` are the top-2 documents returned by this function for the article `Entertainment`.

For sanity check purposes, we also considered the RND function which randomly returns articles. Note that the LKI() and LKO() functions were obtained with the Wikipedia Miner toolkit (Milne and Witten, 2013).

2.3 Explicit Semantic Analysis (ESA-B)

We also implemented the approach described in (Bouamor, 2014) which has been shown by the author to be more accurate than the aforementioned standard approach. The proposed method is an adaptation of the Explicit Semantic Analysis approach described in (Gabrilovich and Markovitch, 2007).

A term to translate is represented by the titles of the Wikipedia articles in which it appears. The projection of the resulting context vector into the target language is obtained by following the available inter-language links.[13] The words of the articles reached this way are candidates to the translation and are further ranked by a tf-idf schema. This approach avoids the need for a seed bilingual lexicon, but uses instead the structure of Wikipedia, and its multilingualism more particularly.

One meta-parameter of this approach is the maximum size of the context vector, that is, the maximum number of article titles to keep for describing a term. One might think that considering all the articles in which a term to translate is found is a good idea, but this strategy faces some sort of *semantic drift*. For instance, while translating the term `tears`, the context vector is populated with articles related to music albums that contain this term in their text content, while the associated French article (when available) almost never contains the translation. We investigate this meta-parameter in section 4. The other parameters were set as recommended in (Bouamor, 2014).

[10] Ergane (12 914 entries - `http://download.travlang.com`), Freelang (38 869 entries - `http://www.freelang.net`), as well as an in-house lexicon (99 747 entries).

[11] Actually, the authors reported that with the LLR association measure, the Dice similarity was a better choice, but we kept along with the cosine measure for simplicity.

[12] `http://www.lucene.org`

[13] Articles with no inter-language links are simply ignored.

3 Experimental Protocol

3.1 Comparable corpus

DBpedia is extracted from Wikipedia (Jens Lehmann, 2014). Thus, we downloaded the Wikipedia dump of June 2013 in both English and French. The English dump contains 4 262 946 articles, and the French one contains 1 398 932. Although some articles that share an inter-language link are parallel (Patry and Langlais, 2011), most article pairs are actually only comparable (Hovy et al., 2013).

3.2 English terms without translation

The vast majority (82,3%) of articles in the English Wikipedia do not have a link to an article in the French Wikipedia. We are interested to identify the translation of their title. Yet, we noticed that many of them are actually describing named entities (persons, geographic places, etc.), which typically do not require translation.[14] In order to filter named entities, we applied the BabelNet filter.[15] We ended up with a list of 521 895 (18,5%) terms we ultimately seek to translate. In this study, we further narrowed down our interest on unigrams.[16] This represents roughly 30% of those English terms.

3.3 Reference List

To evaluate our different approaches, we build a test set — a list of English source terms and their reference (French) translation. For this, we randomly sampled pairs of articles in Wikipedia that are inter-language linked. It is accepted that the titles of a pair of articles inter-language linked often constitute good translations (Hovy et al., 2013). Therefore, for each term (title) of our test set, we collected the associated title as a reference translation.

The sampling was done without considering named entities. For this purpose, we only considered article pairs which English title belongs to the bilingual lexicon we used as a seed lexicon for the STAND approach. Since the frequency of a source term is a key parameter of projective approaches, we also paid attention to vary the frequency range of the English terms we considered in our test set. More precisely, we gathered terms in those different ranges: infrequent [1-25], moderate [26-100], large [101-1000] and huge [1001+], where the frequency is the one in (English) Wikipedia. Some examples of pairs in each range are displayed in Table 2.

[1-25]	74 (8.5%)	
	myringotomy	paracentèse
[26-100]	267 (30.7%)	
	syllabification	césure
[101-1000]	259 (29.8%)	
	numerology	numérologie
[1001+]	269 (30.9%)	
	entertainment	divertissement
Total	869 (100%)	

Table 2: Distribution of the number of test forms at a given frequency range along with an example of an English term and its reference (French) translation.

We measured that using a large parallel corpus,[17] we could only identify the translation of roughly 1% of those terms, which indicates that parallel data might be of little interest in identifying the translations of Wikipedia article titles.

3.4 Evaluation

Our approaches have been configured to produce a ranked list of (at most) 20 candidates for each source (English) term. We compute two metrics to compare them: precision at rank 1 (P@1) which indicates the percentage of terms for which the best ranked candidate is the reference one, and Mean Average Precision at rank 20 (MAP-20), a measure commonly used in information retrieval (Manning et al., 2008) which averages precision at various recall rates.

3.5 Technical considerations

The standard approach (STAND) can be rather computation and time consuming, since any target word in Wikipedia is a potential candidate for

[14] Some languages do involve transliteration, but this is definitely beyond the scope of this paper.

[15] We used the `BabelSynset.getSynsetType()` function of the BabelNet API for this purpose.

[16] Methods that handle multi-word expressions typically embed single word translation (Morin and Daille, 2009); therefore our choice.

[17] We gathered 32 millions of sentence pairs from different available parallel corpora, including the GIGAWORD corpus we downloaded from `http://www.statmt.org/wmt13/translation-task.html`.

a given source term, and we are dealing with a rather large comparable corpus. Just as an illustration, the word `france` occurs more than 1 million times in the French Wikipedia, and its context vector potentially contains as much as 136 514 words (considering a context window of 6 words). Therefore, in our experiments, we only consider the first 50 000 occurrences of each term while populating the context vectors. Also, comparing source and target vectors can be time consuming, especially with context vectors of very high dimension. To save some time (and memory), we only represent a context vector (source or target) by (at most) the 1000 top-ranked terms according to the association measure being used.

4 Results

4.1 STAND

In some calibration experiments,[18] we observed that increasing the size of the window in which we collect the context words leads to noise (see Table 3). The optimal window size was 6 (3 words on each side of the word under consideration, excluding function words), which means that the co-occurrent words should be taken in the immediate vicinity of the term to translate. This corroborates the study in (Bullinaria and Levy, 2007). Therefore, we set the value of this meta-parameter to 6 in the remainder.

|window|	MAP-20
2	0.72
6	0.75
14	0.62
30	0.55

Table 3: MAP-20 of STAND (ORD) measured on a development set, as a function of the window size (counted in word).

The results of two variants of the standard approach are reported in Table 4 (line 1 and 2). Clearly, using ORD as an association measure drastically improves performance. This definitely corroborates the findings of Laroche and Langlais (2010). Still, the differences between both variants is surprisingly high: ORD delivers over six time higher performance than LLR does on average, while in the aforementioned work, the difference was much less marked.[19] Therefore, we use this association measure in the neighbourhood variants we tested.

We observed in practice the tendency of ORD to reward word pairs that appear often together even though the frequency of each word is very low. Thus, the context vector gathered with ORD tend to contain rare words that only appear in the context of the article under consideration. Those words offer a good discriminative power in our task, thus leading to much higher performance than the context vectors computed by LLR, which tend to gather more general related terms. This tendency can be observed in Figure 1 where ORD leads to a context vector with much more specific words. This observation deserves further investigations.

ORD	LLR
myringoplasty (16.32)	**tube** (147.6)
myringa (16.14)	**laser** (44.90)
laryngotracheal (15.13)	**procedure** (40.83)
tympanostomy (14.60)	usually (31.86)
laryngomalacia (14.19)	**knife** (30.13)
patency (13.43)	**myringoplasty** (29.85)
equalized (11.75)	**ear** (28.19)
grommet (11.58)	**laryngotracheal** (27.45)
obstructive (11.09)	**tympanostomy** (26.39)
incision (10.37)	cold (24.09)

Figure 1: Top words in the context vector computed with ORD and LLR for the source term `Myringotomy`. Words in bold appear in both context vectors.

A second observation that can be made is the strong correlation between the frequency of the term to translate and the performance of the approach. As a matter of fact, the performance for very frequent terms ([1001+]) is more than ten times the one measured on infrequent ones ([1–25]). This is a well-know fact that has been analyzed for instance in (Prochasson and Fung, 2011) where the authors report a precision of 60% for frequent test words (words seen at least 400 times), but only 5% for rare words (seen less than 15 times).

Overall, and even if a close comparison is difficult, the results we obtained for STAND are in-

[18] We used a development set of 125 (unigram) terms, considering a candidate list of 50k words randomly selected to which we added the reference translations.

[19] In Table 3 of their article, the authors measured on a test-set of 500 terms a MAP of 0.536 for ORD, and 0.413 for LLR.

	[1-25]		[26-100]		[101-1000]		[1001+]		[Total]	
	P@1	MAP	P@1	MAP	P@1	MAP	P@1	MAP	P@1	MAP
STAND (LLR)	0.000	0.003	0.011	0.019	0.019	0.023	0.134	0.154	0.051	0.061
STAND (ORD)	0.027	0.057	0.217	0.281	0.425	0.474	0.461	0.506	0.338	0.389
STAND (○-100)	0.027	0.058	0.146	0.201	0.154	0.219	0.104	0.162	0.125	0.182
LKI-1000	0.000	0.002	0.064	0.080	0.124	0.156	0.126	0.155	0.096	0.119
LKO-1000	0.000	0.000	0.016	0.022	0.089	0.119	0.033	0.046	0.044	0.058
CMP-1000	0.016	0.022	0.072	0.099	0.131	0.170	0.093	0.120	0.092	0.121
RND-1000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ESA-B	0.014	0.080	0.056	0.122	0.205	0.300	0.424	0.513	0.211	0.293

Table 4: Precision (at rank 1) and MAP-20 of some variants we tested. Each neighbourhood function was asked to return (at most) 1000 English articles. The ESA-B variant is making use of context vectors of (at most) 30 titles.

line with those reported in (Laroche and Langlais, 2010) that also focused on Wikipedia, but mining translations of medical terms. The authors reported a precision at rank one ranging from 20.7% up to 42.3% depending on test sets and configurations considered.

As we discussed in Section 3.5, due to computational issues, we cut the context vectors of the STAND approach after 1 000 terms. In order to measure how sensitive this cut-off is, we computed a variant where the top-100 terms only are kept (considering the association measure). The results of this variant are reported in line 3 of Table 4. As expected, the performance of the STAND approach drops significantly on average, and especially for very frequent terms ([1001+]).

4.2 Neighbourhood variants

We tested our neighbourhood functions as well as several combinations of them. One meta-parameter we investigated is the maximum number of articles returned by a function. We early observed that the more the better, something we explain shortly. Thereafter, each function was asked to return at most 1 000 articles. The results obtained by the 3 neighbourhood functions we described in section 2 are reported in lines 4 to 6 of Table 4.

Clearly, all the neighbourhood variants we considered yielded a significant drop in performance, which is disappointing from a practical point of view. This suggests that there is no obvious way to reduce the number of source documents to consider while populating the context vector of the term to translate. One explanation for this is that in our implementation, the context vector of each target candidate is computed by considering the full (French) Wikipedia collection. This dissymmetry introduces a mismatch between the source and target context vectors, leading to poor performances. A solution to this problem consists in computing target context vectors online from a subset of target documents of interest.[20] A drawback of this solution is (of course) that the computation must take place for each term to translate. This is left as a future work.

At least, the neighbourhood variants we experimented outperform the one where random documents are sampled (RND). This latter variant could not translate a single term of the test set.

4.3 ESA-B

In the default configuration of the approach described in (Bouamor et al., 2013), the authors limit the size of the context vector to 100, which we found suboptimal in our case. We varied the dimension of the context vectors and observed the best value to be 30 (see Table 5). This is the value used in the sequel.

|context|	MAP-20
10	0.248
20	0.287
30	0.293
50	0.291
100	0.271

Table 5: MAP-20 of ESA-B measured on the test set, as a function of the context vector dimension.

[20] This subset could, for instance, be defined by following the inter-language links of the source documents returned by the neighbourhood function.

Somehow contrary to what has been observed in (Bouamor et al., 2013), we observe that ESA-B (P@1 = 0.211) under-performs the STAND approach with the ORD association measure (P@1 = 0.338). One explanation for the difference is that, in (Bouamor et al., 2013), the authors filter in words such as nouns, verbs and adjectives when populating the context vectors, while we do not. This filter might interfere with the observation made in section 4.1 that, with ORD, rare words (which might be filtered out, such as URLs or even spelling mistakes) tend to appear in the context vectors, and happen to help in discriminating translations.

4.4 Analysis

If we consider the 528 test terms that appear over a hundred times in Wikipedia ([101+]), a test case where both approaches perform well, STAND (ORD) translates correctly 362 of them (considering the top-20 solutions), while ESA-B translates 351. If we had an oracle telling us which variant to trust for a given term, we could translate correctly 431 terms (81.6%), which indicates the complementarity of both approaches.

We analyzed the 97 terms for which our two approaches failed to propose the reference translation in the top-20 candidates and we identified a number of recurrent cases we describe hereafter.

First, English terms do appear in the French Wikipedia material that eventually get selected by the STAND approach. This is, for instance the case for the term `barber` (oracle translation: `coiffeur`) for which STAND proposed the translation `barber`.

Second, we observed that STAND (and perhaps ESA-B in a less systematic way) often proposes morphological variants of the reference translation. For instance, `coudre` (a verbal form) is the first proposed translation for `sewing`, while the reference translation is the noun `couture`.

Third, it happens in a few cases that the reference translation, although correct is very specific. Of course this penalizes equally both approaches we tested. For instance, the reference translation of `veneration` is `dulie`, while the first translation produced by STAND is `vénération` (a correct translation).

Also, and by far the most frequent case, we observed a *thesaurus effect* of both approaches where terms related to the source one are proposed. This effect can be observed in Figure 2 in which top candidates proposed by several variants we tested are reported for the terms exemplified in Table 2.

Finally, it happens that the top-20 candidates proposed are just noise (e.g. `noun` translated as `spora`).

5 Discussion

In this study, we implemented and compared two projective approaches for identifying the translation of terms that correspond to articles in English Wikipedia that do not have an inter-language link to an article in the French Wikipedia. Doing so would potentially help in enriching the `rdfs:label` property attached to concepts in `DBpedia`, thus easing semantic annotation in French. One method is a variant of the popular approach pioneered by (Rapp, 1995) which uses a bilingual seed lexicon for mapping source and target context vectors, and the other one has been proposed in (Bouamor et al., 2013) for which the authors shown to deliver state-of-the-art performance.

Among other things, our experiments suggest that the STAND approach performs as well or better than the ESA-B approach and combining both approaches, especially for high frequency terms might improve our results.

We also observed the well-known bias of those approaches toward frequent terms, which urges the need for methods adapted to less frequent terms. As a future work, we will investigate the solution proposed in (Prochasson and Fung, 2011) which is one step in this direction.

Also, the projective methods we considered embed several meta-parameters which values are sensible. It is therefore difficult to know a priori which configuration to chose for a given task, without conducting costly calibration experiments. Having at our disposal a number of different test cases would help in developing expertise in doing so. With the hope that this might help, the code and resources used in this work will be available at this url: `http://rali.iro.umontreal.ca/rali/?q=fr/Ressources`

Acknowledgments

This work has been funded by the Quebec funding agency *Fonds de Recherche Nature et Technologies* (FRQNT).

myringotomy [1-25]

ESA-B	–	laryngologie (0.209) oto (0.191) rhino (0.180) traitement (0.125) otite (0.080)
STAND (ORD)	–	permette (0.0489) devra (0.0473) scopie (0.0471) nécessitait (0.046) pût (0.045)
STAND (LLR)	–	melanosporum (0.274) neural (0.272) séminifère (0.269) ncathodique (0.269)

syllabification [26-100]

ESA-B	–	langues (0.517) consonne (0.420) langue (0.353) lettre (0.223) phonétique (0.166)
STAND (ORD)	–	modifier (0.079) suffit (0.074) vouloir (0.074) syllabique (0.074) intonation (0.072)
STAND (LLR)	–	édicté (0.106) exécutoire (0.097) syllabique (0.096) irrévocable (0.092)

numerology [101-1000]

ESA-B	20	œuvre (0.053) gematria (0.037) angels (0.031) nombres (0.029) chiffre (0.027)
STAND (ORD)	1	numérologie (0.095) occultisme (0.062) ésotérisme (0.062) divinatoire (0.058)
STAND (LLR)	5	jyotish (0.415) conditionaliste (0.412) karmique (0.364) domification (0.358)

entertainment [1001+]

ESA-B	2	entertainment (0.392) divertissement (0.151) vidéo (0.121) sony (0.111) jeu (0.073)
STAND (ORD)	–	beatmakers (0.012) manglobe (0.011) spycraft (0.011) déduplication (0.010)
STAND (LLR)	–	dsi (0.299) eshop (0.294) cocoto (0.231) ead (0.225) imagesoft (0.210)

Figure 2: Top candidates produced by several variants of interest for some test terms. The second column indicates the rank of the oracle translation when present in the top-20 returned list (or – if absent).

References

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22.

Dhouha Bouamor, Adrian Popescu, Nasredine Semmar, and Pierre Zweigenbaum. 2013. Building specialized bilingual lexicons using large scale background knowledge. In *EMNLP*, pages 479–489.

Dhouha Bouamor. 2014. *Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables*. PhD thesis, Université Paris Sud - Paris XI, February.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526, August.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Comput. Linguist.*, 19(1):61–74, March.

Stefan Evert. 2005. *The statistics of word cooccurrences*. Ph.D. thesis, Dissertation, Stuttgart University.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, AMTA '98, pages 1–17, London, UK, UK. Springer-Verlag.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John McCrae. 2012. Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:63–71, March.

Asunción Gómez-Pérez, Daniel Vila-Suero, Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado-de Cea. 2013. Guidelines for multilingual linked data. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, pages 3:1–3:12, New York, NY, USA. ACM.

Amir Hazem, Morin Emmanuel, and others. 2013.

Word co-occurrence counts prediction for bilingual terminology extraction from comparable corpora. *IJCNLP 2013*.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27, January.

Robert Isele Jens Lehmann. 2014. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 617–625, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

David Milne and Ian H. Witten. 2013. An open-source toolkit for mining wikipedia. *Artif. Intell.*, 194:222–239, January.

Emmanuel Morin and Béatrice Daille. 2009. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, page 0, August.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, December.

Alexandre Patry and Philippe Langlais. 2011. Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, pages 87–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emmanuel Prochasson and Pascale Fung. 2011. Rare word translation extraction from aligned comparable documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1327–1335, Stroudsburg, PA, USA. Association for Computational Linguistics.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, pages 320–322, Stroudsburg, PA, USA. Association for Computational Linguistics.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2013. Overviewing important aspects of the last twenty years of research in comparable corpora. In Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, and Pascale Fung, editors, *Building and Using Comparable Corpora*, pages 1–17. Springer Berlin Heidelberg.