Merging knowledge bases in different languages

Jerónimo Hernández-González
University of the Basque Country, Donostia, GI, Spain
jeronimo.hernandez@ehu.eus

Estevam R. Hruschka Jr.
Federal University of Sao Carlos, São Carlos, SP, Brazil
estevam@dc.ufscar.br

Tom M. Mitchell
Carnegie Mellon University, Pittsburgh, PA, USA
tom.mitchell@cs.cmu.edu

Abstract

Recently, different systems which learn to populate and extend a knowledge base (KB) from the web in different languages have been presented. Although a large set of concepts should be learnt independently from the language used to read, there are facts which are expected to be more easily gathered in local language (e.g., culture or geography). A system that merges KBs learnt in different languages will benefit from the complementary information as long as common beliefs are identified, as well as from redundancy present in web pages written in different languages. In this paper, we deal with the problem of identifying equivalent beliefs (or concepts) across language specific KBs, assuming that they share the same ontology of categories and relations. In a case study with two KBs independently learnt from different inputs, namely web pages written in English and web pages written in Portuguese respectively, we report on the results of two methodologies: an approach based on personalized PageRank and an inference technique to find out common relevant paths through the KBs. The proposed inference technique efficiently identifies relevant paths, outperforming the baseline (a dictionary-based classifier) in the vast majority of tested categories.

1 Introduction

In the last few decades, the machine learning community has launched different research projects to take advantage of the massive source of information which has become the web, and of the people who build it up. Among others, information extraction systems (IES) which use the text found in webpages to extract, validate and incorporate beliefs to a structured knowledge base have been developed (e.g., YAGO (Suchanek et al., 2008), NELL (Mitchell et al., 2015) or Knowledge Vault (Dong et al., 2014)). Such knowledge bases (KBs) store facts about the real world, which are represented as entities and relationship among entities. The reliability of a fact, inferred from the web, is at first questionable due to the *noisy* information available on the Web. This difficulty is usually overcome by relying on data redundancy from multiple Web pages. Requiring higher degrees of redundancy to incorporate beliefs to the KB help to improve the quality of the learnt KB.

In the long run, having two such IES, running independently, tend to generate equivalent KBs, even if they gather information from different webpages, in different time or using different terminology. However, if those two systems extract (and store) facts from Web pages written in different languages, it can be hard to automatically identify redundant facts, or to automatically merge such KBs. Let us assume an English KB containing the concept of the city of *Sao Carlos* as a belief and, also, a Portuguese KB with the equivalent concept represented in Portuguese as *São Carlos*. Let us assume also that, in the Portuguese KB, *São Carlos* is linked to the concept *Paulo Altomani*, the major of the city. Combining both KBs and identifying the equivalence between *Sao Carlos* (en) and *São Carlos* (pt) can help to automatically populate the English KB adding the fact that *Paulo Altomani* is the mayor of *Sao Carlos* (en).

In this paper we deal with the problem of merging KBs learnt in different languages. This task consists of ontology alignment and equivalent concept matching. We use graph-based inference techniques to deal with the problem of identifying

Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, ACL 2017, pages 21–29,
Vancouver, Canada, August 3, 2017.

equivalent entities in different KBs assuming that, in spite of being learnt independently, they share a common ontology. The contributions of this work are as follows:

- An approach to multi-lingual KB merging based on Personalized PageRank
- A path-based graph inference approach that shows promising results when compared to Personalized PageRank.
- An empirical analysis by means of a case study. When graph connectivity is enhanced by means of a SVO corpus, results stand out.

In the remainder of this paper we first provide a formal description of the problem, which is formulated as an inference problem. Then, the proposed solutions are presented. In Section 4, the different approaches are tested in a case study with two KBs independently learnt by NELL (Mitchell et al., 2015) from English and Portuguese webpages respectively. Next, their behavior is discussed. The paper finishes with conclusions and ideas for future work.

2 Framework

Consider a Knowledge Base (KB) K as a tuple (O, I_c, I_r, S). The ontology O is represented by a 4-tuple (C, H_C, R, H_R), where H_C codifies the hierarchy among categories $c \in C$ (e.g., the category *city* is a specification of the category *place*) and, similarly, H_R codifies the hierarchy among relation-types $r(c_1, c_2) \in R$ with $c_1, c_2 \in C$ (e.g., *locatedAt*(*city*, *country*) is more general than *capitalOf*(*city*, *country*)). I_c and I_r are the sets of entities and relationships, respectively, that populate the KB. Thus, an instance $(e, c) \in I_c$ assigns entity e to category $c \in C$ (e.g., (*Pittsburgh*, *city*) or (*USA*, *country*)) and each instance $(e_1, r, e_2) \in I_r$ is a relation of type r which associates two entities, e_1 and e_2, with $(e_1, c_1) \in I_c$, $(e_2, c_2) \in I_c$ and $r(c_1, c_2) \in R$ (e.g., (*Pittsburgh*, *locatedAt*, *USA*)). S involves the literal strings which are used to refer to the entities. Two —or more— literal strings s_1 and s_2 ($s_1 \neq s_2$) can refer to the same entity e ($(s_1, e) \in S \wedge (s_2, e) \in S$), and the same literal string s can refer to two different entities e_1 and e_2 as well ($(s, e_1) \in S \wedge (s, e_2) \in S$ with $e_1 \neq e_2$). For example, the literal strings "Steal City" and "Pittsburgh" can refer to the concept *Pittsburgh*, whereas "New York" could refer to both *NYC* and *New York State*.

Graph representation. In this paper, a graph representation of the KB is used and the problem of merging KBs in different languages is handled as a graph inference problem. Each entity e is represented as a node. Each relation instance $(e_1, r, e_2) \in I_r$ is represented by an edge of type r between the nodes representing entities e_1 and e_2. Similarly, for each $(s, e) \in S$, string s is represented as a node and an edge of type *canReferTo* links it to entity e. In the remaining of the paper, the terms "nodes" and "entities", on the one hand, and "relation" and "edge types", on the other, are interchangeably used.

2.1 Inferring entity-equivalence across KBs in different languages

For the sake of simplicity, let us follow the example of our case study to describe the problem of merging two KBs which share the same ontology structure (categories and relation types) but which have been learnt (populated) in different languages. Given both KBs, $K^{en} = (O, I_c^{en}, I_r^{en})$ and $K^{pt} = (O, I_c^{pt}, I_r^{pt})$, in English and Portuguese respectively, the merging process $K^* = merge(K^{en}, K^{pt})$ consists mainly of the union of both sets of entities $I_c^* = I_c^{pt} \cup I_c^{en}$, where only an instance of the equivalent entities across languages (e.g., *New York* or *Nova Iorque*) remains. As a consequence of this first step, the sets of relation instances $I_r^* = I_r^{pt} \cup I_r^{en}$ is similarly merged: two relation instances in different languages, (e_1^{en}, r, e_2^{en}) and (e_1^{pt}, r, e_2^{pt}), are equivalent if their relation type $r \in R$ is the same and the associated entities are fused in I_c^* ($e_1^{en} \sim e_1^{pt}$ and $e_2^{en} \sim e_2^{pt}$). To avoid losing information, per-language literal string sets (S^{en}, S^{pt}) are kept linked to the corresponding entities in I_c^*.

The key step is, therefore, the identification of equivalent entities across languages. Let us introduce the relation types $(e^{en}$, *equivalentTo*, $e^{pt})$, which connects two entities in different language specific KBs, and $(s^{en}$, *canBeTranslatedAs*, $s^{pt})$, which relates two literal strings which are the translation of each other in the different languages. Thus, the originally independent language specific KBs become connected and the problem of finding equivalent entities across languages can be reformulated as inferring the existence of *equivalentTo* relationships (edges) between pairs of entities (nodes) in different KBs (subgraphs).

3 Methods

In this study, we make use of inference techniques over graphs to find equivalent entities in different language KBs: a Personalized PageRank (PPR) (Haveliwala, 2002) based approach and another one based on Path Ranking Algorithm (PRA) (Lao and Cohen, 2010). Both techniques produce classification models which, given a new pair of entities, predict whether or not an *equivalentTo* relationship is suitable among them.

3.1 Personalized PageRank based approach

In the context of webpage ranking, Personalized PageRank (Haveliwala, 2002) was designed to bias the result of the original PageRank algorithm (Page et al., 1999) to make it topic-sensitive. It can be seen as a similarity measure that characterizes the neighborhood of a node X in a graph. Formally, it estimates a probability distribution over the nodes of the graph. Considering X as the source node of a random walk, it estimates the probability of reaching node Y after w random steps. At time t, the next step follows one of the out-edges of current node X_t with equal probability $(1-\alpha)\cdot\frac{1}{|X_t|}$ (where $|X_t|$ is the out-degree of node X_t) or jumps back to the source node X with probability α. The stationary probability distribution is usually approximated by sampling a number n of random walks with probability α of restarting at source node X. The probability assigned to node Y is the proportion of walks which finish at Y.

In the context of this work, PPR has been used to measure similarity between nodes. Assuming that two equivalent entities in different language subgraphs (L_1 and L_2) will be highly connected through a number of different paths, the equivalent entity $e_t^{L_2}$ is expected to be assigned a high probability by a PPR with origin at entity $e_o^{L_1}$. Using PPR to estimate the probability distribution $p(\cdot|e_o^{L_1})$, a classification model is built by imposing three conditions: (1) the predicted equivalent entity belongs to a different language subgraph ($e_t^{L_2} : L_1 \neq L_2$), (2) the category of both the source and target entities is the same or compatible (both are in the same hierarchical line in H_C):

$$(e_o^{L_1}, c_o) \in I_c^{L_1} \wedge (e_t^{L_2}, c_t) \in I_c^{L_2} :$$

$$c_o, c_t \in C \wedge (c_o = c_t \vee c_o \overset{H_C}{\rightleftharpoons} c_t)$$

and (3) the probability of the predicted entity exceeds threshold $h \leq p(e_t^{L_2}|e_o^{L_1})$.

3.2 Path Ranking algorithm based approach

The Path Ranking algorithm (Lao and Cohen, 2010) transforms the task of inferring new relationships of type r between pairs of entities into a binary classification problem: given a new pair of nodes, is a relationship of type r suitable between them? To do so, it generates, in two steps, a training matrix from which any type of classifier can be learnt. The pairs of nodes already connected by a relationship of type r are positive pairs or examples in this approach. During the first step, paths (sequence of relation types, $r_1, r_2, \ldots, r_p$) commonly connecting the nodes of the positive pairs are identified by running a number of random walks of limited length. In the second step, a training matrix is built such that each identified path constitutes a feature (column) and each pair is a positive example (row). Each cell (i, j) of the matrix is assigned the probability of reaching the target node e_t^i of the i-th pair using a random walk that follows the sequence of relation types of the j-th path with origin at node e_o^i.

Departing from the original design, the generation of paths has been adapted to take advantage of the particularities of our application. First of all, note that every path which connects two nodes in different language subgraphs, e^{en} and e^{pt}, includes an *equivalentTo* or *canBeTranslatedAs* relation type. Note also that, assuming a common ontology for both KBs, the relation types and categories are the same in both languages. The idea behind the original PRA —i.e., certain relationships (or paths) can be particularly relevant for determining the equivalence of entities of a specific category— is extended to the multi-language context by looking for relevant paths which appear replicated in both language subgraphs. Suppose that there is a *Di Blassio* entity in both languages and an *equivalentTo* relationship links them. Suppose also that there are (*Di Blassio*, *isMayorOf*, *New York City*) and (*Di Blassio*, *isPrefeitoDe*, *Cidade de Nova Iorque*) relationships in English and Portuguese subgraphs respectively. Knowing that (*New York City*, *Cidade de Nova Iorque*) are equivalent, *isMayorOf* can be considered as a relevant path (of length one) to predict the equivalence of *cities*. Intuitively, a person cannot be mayor of different cities.

Our search for relevant paths starts, for each positive pair, with a breadth-first search from the source node $e_o^{L_1}$ looking for nodes $e_b^{L_1}$ with an across-language edge (edge type *equivalentTo* or *canBeTranslatedAs*). For each of these nodes, the node $e_b^{L_2}$ at the other extreme of the across-language relationship is taken. Then, the path followed to reach $e_b^{L_1}$ from $e_o^{L_1}$ is reversed. If the reversed path connects $e_b^{L_2}$ to the target node $e_t^{L_2}$ of the corresponding positive pair, the whole path from $e_o^{L_1}$ to $e_t^{L_2}$ is kept for evaluation. The relevance of a path $(r_1, r_2, \ldots, r_p)$ is measured as the probability of reaching the target node following a random walk (through relationships of types $r_1, r_2, \ldots$) starting at source node, or in the opposite direction. Thus, the most relevant path always leads to the opposite node of the pair, and only to it. Uninformative paths, those whose rates are below the average, are filtered out. With the remaining relevant paths, a training matrix is built in the same way as the original PRA.

4 Experiments

A complete set of experiments has been designed to test the performance of both approaches in the task of identifying equivalent concepts across languages. A baseline based on dictionary translations is used to put these results in context.

4.1 Knowledge bases

The knowledge bases used in these experiments correspond to the 970th and 110th iterations of the English and Portuguese versions of NELL, respectively. As aforementioned, a graph is obtained from each KB drawing a node for each entity and literal string and a labeled edge for each relationship among entities. Moreover, edges of type *canReferTo* link each entity with its literal strings. We found out that many entities in both KBs are isolated, i.e., they have no relationship. For these experiments, all the isolated nodes have been pruned from the graphs as our techniques cannot deal with them: both presented techniques make use of the relationships among entities to perform. The resulting graph is used below in a first set of experiments.

As previously mentioned, both PPR and PRA-based techniques make use of relationships and, in fact, they need well connected graphs to perform correctly. However, the graphs obtained from NELL KBs are quite sparse (see Table 1 for its mean out-degree). An enhanced connectivity among entities is achieved considering a SVO corpus. A SVO consists of statistics about the presence of a triplet *subject-verb-object* in a text corpus usually crawled from the Web. In this study, Wijaya and Mitchell (2016) method to map verbs found in a corpus to relationships of a given structured KB has been used. It explores a SVO corpus looking for verbs which can be used to represent the different relation types $r \in R$ of an ontology O. Given the returned set of representative verbs for a specific relation type r, pairs of literal strings $s_1, s_2 \in S$ which appear linked by means of one or more representative verbs in the SVO corpus can be considered as evidence of a r relationship. In practice, all the entities which can be referred to by s_1 and s_2 are connected by means of an edge of type r. As can be observed in Table 1, connectivity is largely enhanced. On average, the number of edges connecting each node has increased although, according to the related standard deviations, the behavior is not uniform. The graph resulting from this enhancing process is used in a second set of experiments.

Note that the enhancement with SVO-inferred

	English	
Category	GRAPH 1	GRAPH 2
animal	13.32 ± 47.12	392.02 ± 1859.43
country	22.65 ± 68.60	289.97 ± 970.55
city	5.00 ± 22.31	134.72 ± 1176.05
movie	1.60 ± 1.69	88.95 ± 891.37
person	2.99 ± 6.76	244.10 ± 1394.07
writer	2.72 ± 4.75	43.50 ± 419.41
actor	2.19 ± 2.12	77.21 ± 819.10
sport	19.94 ± 132.71	256.49 ± 1430.06
all	4.48 ± 28.95	275.49 ± 1715.53
	Portuguese	
Category	GRAPH 1	GRAPH 2
animal	1.57 ± 1.17	27.23 ± 58.58
pais	5.04 ± 15.99	180.79 ± 820.71
cidade	1.71 ± 4.33	32.03 ± 135.13
filme	1.33 ± 0.79	34.77 ± 177.64
pessoa	1.18 ± 0.66	16.62 ± 101.67
escritor	1.15 ± 0.46	6.40 ± 12.04
ator	1.67 ± 1.41	6.96 ± 11.40
esporte	3.88 ± 6.31	26.98 ± 83.30
all	1.81 ± 3.93	51.26 ± 199.03

Table 1: For each language and category, mean out-degree value and associated standard deviation of the nodes of that category in the (first) graph, without isolated nodes, and in the (second) graph, fed with SVO-inferred relationships before pruning. The last row sums up all the categories.

	English					
Category	UNPROCESSED		GRAPH 1		GRAPH 2	
animal	12,436	(36)	591	(23)	746	(27)
country	6,031	(106)	443	(93)	460	(93)
city	18,893	(460)	4,437	(237)	5,311	(263)
movie	7,008	(42)	712	(38)	831	(38)
person	6,693	(403)	2,898	(395)	3,050	(399)
writer	18,911	(61)	1,707	(39)	2,143	(40)
actor	28,361	(512)	794	(139)	1,421	(167)
sport	5,022	(109)	205	(66)	381	(75)
all	1,909,339	(4,126)	66,239	(2,112)	96,086	(2,331)

	Portuguese					
Category	UNPROCESSED		GRAPH 1		GRAPH 2	
animal	101	(36)	63	(14)	97	(35)
pais	153	(106)	94	(87)	136	(103)
cidade	5,767	(460)	483	(138)	1,404	(282)
filme	368	(42)	64	(25)	132	(40)
pessoa	621	(403)	611	(304)	614	(376)
escritor	114	(61)	26	(23)	63	(37)
ator	1,870	(512)	129	(36)	793	(208)
esporte	153	(109)	34	(29)	125	(94)
all	30,401	(4,126)	5,119	(1,827)	12,930	(2,565)

Table 2: For each language subgraph and category, the number of entities and, from these, the number of entities contained in a positive pair are shown. The three columns show counts, from left to right, for (1) the unprocessed graph, (2) the first graph, without isolated nodes, and (3) the second graph, fed with SVO-inferred relationships before pruning. The last row sums up all the categories.

relationships reduces the number of isolated nodes and, therefore, the number of pruned entities decreases. Table 2 reflects the effect of this enhancement, in terms of the number of remaining entities, on the pruning process. In the case of English, the second graph (with SVO-inferred relationships) is almost a 50% larger than the first graph. The Portuguese subgraph, in turn, grows 2.5 times.

4.2 Bridges among both language-specific KBs

Two different strategies have been carried out in this study to generate an initial set of *equivalentTo* relationships. On the one hand, a costly manual introduction of equivalence relationships was carried out. This procedure, although costly, provides highly reliable instances of the relationship. Around 400 fully reliable relationships were thus generated. On the other hand, entities which have the same name (simple matching), in spite of having been learnt in different languages, and belong to compatible categories have been considered as equivalent pairs. Both conditions are fulfilled by up to $4,000$ pairs of entities, among which *equivalentTo* edges have been added. Although the evidence may be strong, this automatically generated set of *equivalentTo* edges could involve misleading information. For example, using this approach a hypothetical entity referring to the renowned machine learning researcher (*Michael Jordan*, *pessoa*) in Portuguese could be connected to the former basketball player (*Michael Jordan*, *athlete*) in English. The pruning process explained above affects these entities too, as shown in Table 2.

Connectivity among language subgraphs at the level of literal strings (relation type *canBeTranslatedAs*) is achieved by means of a dictionary. A list of string translations has been generated combining terms found in WordNet (de Paiva and Rademaker, 2012; Fellbaum, 1998) and translations on demand making use of the Google Translate API [1]. In total, 1.4 million string translations have been obtained. These translations are used to connect nodes representing literal strings in both language subgraphs by means of edges of type *canBeTranslatedAs*.

4.3 Training examples

Standard supervised classification takes advantage of a fully labeled dataset with examples of all the classes. They are necessary to train the classification models as well as to evaluate them. The classification task at hand is a weakly supervised classification problem (Hernández-González et al., 2016); specifically, a positive-unlabeled classification problem (Calvo et al., 2007) where only positive examples are available for training: the pairs of entities related by a *equivalentTo* relationship. No negative example, understood as a pair of nodes in different language subgraphs which are not suitable to hold an *equivalentTo* relationship, is available.

However, in this context, safe procedures for generating negative examples can be figured out. Figure 1 graphically describes the procedure followed in this study, which is based on the assumption that an entity has only one equivalent entity in the opposite language KB. Thus, the source node of a positive pair *is not equivalent* to any node in the opposite subgraph different from the corresponding target node. Formally, each $(e_o^{L_1}$,*equivalentTo*$, e_t^{L_2})$ relationship already present in the KB is individually consid-

[1] https://cloud.google.com/translate/

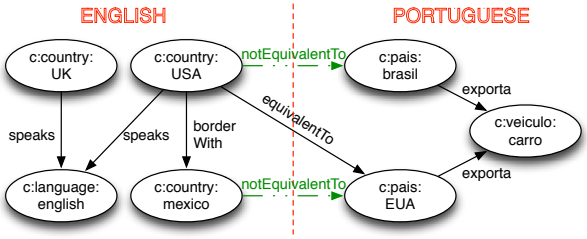

Figure 1: Generation of negative examples: given a positive pair, (*USA*, *EUA*), a walk is launched until an entity with the same category is reached: e.g., (*Mexico*,*Country*) by (*USA*, *borderWith*, *Mexico*). Obtained negative example: (*Mexico*, *EUA*).

ered. For entity $e_t^{L_2}$, other entities $e_{t'}^{L_2}$ with the same or a compatible category are identified in the same language subgraph L_2. A negative example $(e_o^{L_1}$,*notEquivalentTo*$, e_{t'}^{L_2})$ is then built using the original entity $e_o^{L_1}$ and any compatible neighbor $e_{t'}^{L_2}$. For instance, as displayed in Fig. 1, knowing that *USA* is equivalent to the Portuguese *EUA*, *Mexico* is reached from node *USA* through their common relationship (*borderWith*), thus generating a negative example (*Mexico*, *notEquivalentTo*, *EUA*). The same procedure can be carried out in the opposite direction: fixing $e_t^{L_2}$ and looking for compatible entities $e_{o'}^{L_1}$ in the neighborhood of $e_o^{L_1}$. For each positive pair, up to 2 negative pairs among all the negative examples generated in both directions are randomly selected.

4.4 Experimental settings

In addition to both described techniques, a classifier exclusively based on a dictionary is also considered. It predicts an equivalence if the nodes of the query pair represent entities with literal strings which are the translation of each other. Given that a dictionary is probably the simplest solution to deal with this problem, it has been used in this paper as a baseline.

The PPR method has been configured for these experiments with 2,000 random walks of length 5 to estimate the probability distribution, using a probability of restart equal to 0.01. Regarding our PRA-based technique, the breadth-first search is carried out to a depth of 2. These values have been selected within a 20×5-fold cross validation. PPR and the baseline do not need a training step and the results are calculated over the whole training set. For each category, PRA learns a logistic regression classifier (its implementation in Weka (Frank et al., 2016)), which is evaluated in a 10×5-fold cross validation. Remember that positive pairs are *equivalentTo* relationships, which are also represented in the graph. The edges of the graph corresponding to training examples are removed for training and testing.

The PPR and PRA-based approaches, together with the baseline, have been applied over graphs 1 and 2 (without and with SVO-inferred relationships, respectively). As our PRA-based approach learns a classifier per category, a diverse set of eight categories has been selected to test the proposals and report their performance: *animal*, *country*, *city*, *movie*, *person*, *writer*, *actor* and *sport*. In Figure 2, precision-recall (PR) curves are used to describe the results in the first graph. Each subfigure displays the results for one of the selected categories. Following the same layout, Figure 3 shows the results in the second graph. Note that results in figures 2 and 3 are not directly comparable as they have been obtained from training sets of different sizes (see in Table 2 the number of positive entities remaining after pruning in graphs 1 and 2).

5 Discussion

The performance of the different techniques has been assessed for eight categories using two graphs of different sparsity. Results show the competitiveness of the solution based exclusively on a dictionary as well as the outstanding performance of our PRA-based proposal. As expected for an inference technique that intensively explores the graph looking for relevant paths, the use of the more dense SVO+pruned graph enhances the performance of the PRA-based proposal. The behavior of PPR is less regular and changes considerably among categories.

The dictionary connects, across languages, literal strings, which can be used to refer to different entities. This may affect the precision of the dictionary approach: more than one node may be reached following the across-language path *canReferTo*+*canBeTranslatedAs*+*canReferTo* from a single source node. To alleviate this effect, our implementation only predicts a positive equivalence if both nodes of a query pair have the same category. As observed in figures 2 and 3, this baseline reaches precision values equal to 1 for all the categories with the exception of *country* and *person*. Moreover, the size of the dictionary determines the

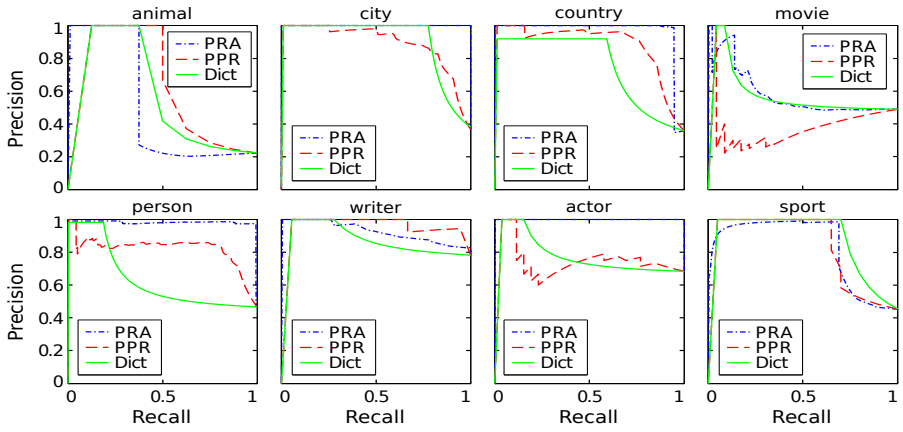

Figure 2: PR curves comparing both proposals with the dictionary as a baseline. Each figure displays the results of the three approaches with the examples of a specific category using the first graph.

maximum recall that this classifier can show before a sharp drop in precision. Thus, this approach is very competitive in categories where our dictionary translates many of the pairs (e.g., *sport*), its performance is limited in categories where few pairs are translated (e.g., *movie*).

The results of the PPR approach are difficult to interpret; no clear pattern is observed. The incorporation of new relationships from the SVO corpus does not enhance its results. Quite the opposite, it seems to harm the results in categories such as *country* or *writer*. This incorporation increases the number of edges among entities of the same language subgraph (see Table 1), while the across language connections remain the same. Intuitively, the probability of a random walk moving across language subgraphs decreases. And this crossing movement is indispensable for the PPR approach to succeed. In categories such as *animal* or *sport*, the behavior of the PPR approach matches that of the dictionary. The short path *canReferTo+canBeTranslatedAs+canReferTo* usually has few instantiations, easily leading from source to target node. Only in a few categories is the PPR approach able to overcome the baseline and, in these cases, the PRA-based approach usually outperforms it. Our PRA-based technique also imitates the dictionary approach. Intuitively, the translation path is usually considered as relevant by this technique. However, even a more complete dictionary would still lack precision in certain cases. According to its unquestionable enhanced performance, our PRA-based technique solves this problem probably relying on both the dictionary and other relevant paths. Specifically, it is able to overcome the baseline when the dictionary is not completely precise (categories *country* and *person*). Finally, only the PRA-based technique clearly improves with the new SVO-inferred edges (categories *animal*, *writer* and *movie*).

A strategy for generating the initial *equivalentTo* relationships (Section 4.2) is the simple matching of entity names in the different languages. This already covers 403 out of 621 entities of category *person* in Portuguese. This rate is lower in category *writer*, although the same behavior would be expected since in both categories proper nouns, which are rarely translated, are used to name entities. There are two possibilities for the remaining entities: they are represented in English with a different name or they do not overlap. We carried out a manual inspection of these entities (219 in the case of *person*) to gain insight, revealing that the majority of them have Portuguese names, and those with English names do not appear in the English KB. Only a few cases have been found where a possible equivalence is present in the KBs with a slightly different name (e.g., *Max Nicholson* in English and *Dr. Max Nicholson* in Portuguese). This observation supports the idea that entities in this kind of categories, which use proper nouns, are usually complementary if an equivalence with exactly the same name is not found. In this context, the recall of the simple matching approach is expected to stand out. Our PRA-based solution would still be competitive in these categories assessing equivalences for entities with slightly different names.

The lower the number of training positive pairs for a category (Tab. 2), the worse the results of

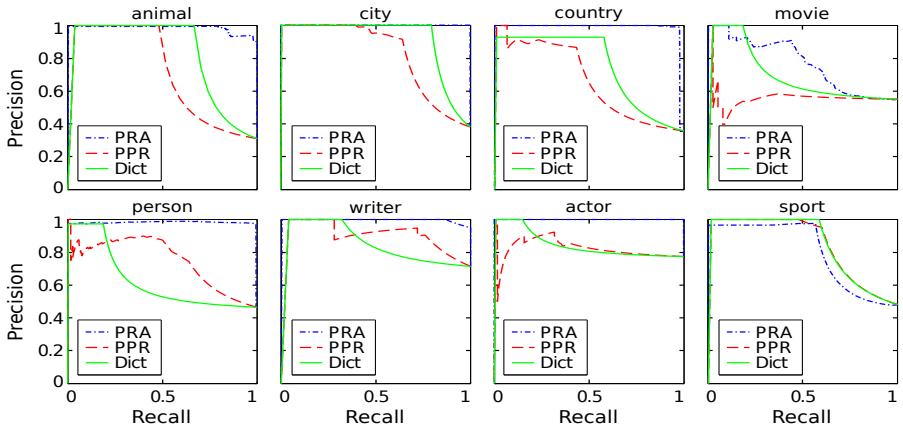

Figure 3: PR curves comparing both proposals with the dictionary as a baseline. Each figure displays the results of the three approaches with the examples of a specific category using the (second) graph pruned after populating it with new relationships inferred from a SVO corpus.

the PRA-based technique (see results for *animal*, *movie*, *writer* and *sport*). The inclusion of the relationships derived from the SVO corpus involves a considerably larger number of positive pairs in all the categories. The performance gain is noteworthy in three of them: *animal*, *movie* and *writer*. However, a larger number of examples does not explain the enhanced performance shown by the PRA-based approach, for example, in the *animal* category. A more densely connected graph is expected to benefit our PRA-based approach, although a larger set of edges does not directly imply a better performance. For example, despite the fact that the category *sport* shows one of the largest out-degree averages (see Tab. 1), the PRA-based classifier can neither overcome the baseline nor the PPR approach (even using the SVO-enlarged graph). Not only does the PRA-based approach require a large number of relationships, it also requires relevant paths. However, it is more likely to find a relevant path in densely connected graphs, such as that obtained after the massive incorporation of relationships from the SVO corpus. That explains the enhanced performance of our PRA-based method in categories *writer*, *animal* and *movie* regarding the results in the first graph without SVO-inferred relationships.

It can be agreed that the larger the number of merged KBs, the more the information which such a multi-lingual system can take advantage of. The approaches proposed in this study are designed to deal with two language subgraphs. However, applying the proposed methodology by means of pairwise comparisons, along with the transitive property, a larger set of equivalent entities will probably be found. For example, if *New York City* is equivalent to *Cidade de Nova Iorque* (pt) and *Cidade de Nova Iorque* (pt) is equivalent to *Ciudad de Nueva York* (es), *New York City* is equivalent to *Ciudad de Nueva York* (es). Whenever the KBs learnt in the different languages are diverse enough —although partial intersection is necessary—, the probability of finding this type of triangulations rises with the number of KBs.

6 Conclusions

In this paper, we deal with the problem of merging two knowledge bases learnt from text written in different languages. Two strategies have been designed and compared with a baseline exclusively based on a dictionary. The proposed solution based on the path ranking algorithm outperforms the baseline and a second proposal based on personalized PageRank.

The PRA-based approach efficiently finds relevant paths between positive pairs of entities. The relevance of a path between two nodes is measured according to the number of entities reached following the path, in both directions. According to the experimental results, it identifies relevant paths in the majority of tested categories, specifically when a more densely connected graph is used.

For future work, taking the KB merging process as a chance for improvement, an approach to co-reference resolution could be to identify entities which have two or more equivalent entities in the opposite language subgraph. The categories of two equivalent entities could also be reassessed if

these are not coincident. If this proposal is integrated into the iterative learning process of NELL, it will benefit from new entities and relationships at each new iteration, possibly leading to the discovery of new relevant paths. Before, as NELL currently allows its ontology to evolve, these proposals should be adapted to deal with unaligned ontologies, similar to what Delli Bovi et al. (2015) or Dutta et al. (2014) do.

Acknowledgments

This work was partially supported by the Basque Government, the Spanish Ministry of Economy and Competitiveness and the University of the Basque Country (IT609-13, Elkartek BID3A, TIN2016-78365-R, University-Society Project 15/19).

References

Borja Calvo, Pedro Larrañaga, and Jose A. Lozano. 2007. Learning Bayesian classifiers from positive and unlabeled examples. *Pattern Recognit. Lett.* 28(16):2375–2384.

Valeria de Paiva and Alexandre Rademaker. 2012. Revisiting a Brazilian WordNet. In *Proc. 6th Global WordNet Conf. (GWC)*. Matsue.

Claudio Delli Bovi, Luis Espinosa-Anke, and Roberto Navigli. 2015. Knowledge base unification via sense embeddings and disambiguation. In *Conf. on Empirical Methods in Natural Language*. ACL, pages 726–736.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. ACM, pages 601–610.

Arnab Dutta, Christian Meilicke, and Simone Paolo Ponzetto. 2014. A probabilistic approach for integrating heterogeneous knowledge sources. In *European Semantic Web Conf.*. Springer, pages 286–301.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, chapter The WEKA Workbench. 4th edition.

Taher H Haveliwala. 2002. Topic-sensitive pagerank. In *Proc. 11th int. World Wide Web Conf.*. ACM, pages 517–526.

Jerónimo Hernández-González, Iñaki Inza, and Jose A. Lozano. 2016. Weak supervision and other non-standard classification problems: a taxonomy. *Pattern Recognit. Lett.* 69:49–55.

Ni Lao and William W. Cohen. 2010. Relational retrieval using a combination of path-constrained random walks. *Mach. Learn.* 81(1):53–67.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-ending learning. In *Proc. 29th AAAI Conf. Artificial Intelligence (AAAI)*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Tech. Report 1999-66, Stanford InfoLab.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. Yago: A large ontology from wikipedia and wordnet. *J. Web Semant.* 6(3):203–217.

D. T. Wijaya and T. Mitchell. 2016. Mapping verbs in different languages to knowledge base relations using web text as interlingua. In *Proc. 15th Annu. Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol. (NAACL)*.