Knowledge Base Unification via Sense Embeddings and Disambiguation

Claudio Delli Bovi
Department of
Computer Science
Sapienza University of Rome
dellibovi@di.uniroma1.it

Luis Espinosa-Anke
Department of Information and
Communication Technologies
Universitat Pompeu Fabra
luis.espinosa@upf.edu

Roberto Navigli
Department of
Computer Science
Sapienza University of Rome
navigli@di.uniroma1.it

Abstract

We present KB-UNIFY, a novel approach for integrating the output of different Open Information Extraction systems into a single unified and fully disambiguated knowledge repository. KB-UNIFY consists of three main steps: (1) disambiguation of relation argument pairs via a sense-based vector representation and a large unified sense inventory; (2) ranking of semantic relations according to their degree of specificity; (3) cross-resource relation alignment and merging based on the semantic similarity of domains and ranges. We tested KB-UNIFY on a set of four heterogeneous knowledge bases, obtaining high-quality results. We discuss and provide evaluations at each stage, and release output and evaluation data for the use and scrutiny of the community[1].

1 Introduction

The breakthrough of the Open Information Extraction (OIE) paradigm opened up a research area where Web-scale unconstrained Information Extraction systems are developed to acquire and formalize large quantities of knowledge. However, while successful, to date most state-of-the-art OIE systems have been developed with their own type inventories, and no portable ontological structure. In fact, OIE systems can be very different in nature. Early approaches (Etzioni et al., 2008; Wu and Weld, 2010; Fader et al., 2011) focused on extracting a large number of relations from massive unstructured corpora, mostly relying on dependencies at the level of surface text. Systems like NELL (Carlson et al., 2010) combine a hand-crafted taxonomy of entities and relations with self-supervised large-scale extraction from the Web, but they require additional processing for linking and integration (Dutta et al., 2014).

More recent work has focused, instead, on deeper language understanding, especially at the level of syntax and semantics (Nakashole et al., 2012; Moro and Navigli, 2013). By leveraging semantic analysis, knowledge gathered from unstructured text can be adequately integrated and used to enrich existing knowledge bases, such as YAGO (Mahdisoltani et al., 2015), FREEBASE (Bollacker et al., 2008) and DBPEDIA (Lehmann et al., 2014). A large amount of reliable structured knowledge is crucial for OIE approaches based on distant supervision (Mintz et al., 2009; Riedel et al., 2010), even when multi-instance multi-learning algorithms (Surdeanu et al., 2012) or matrix factorization techniques (Riedel et al., 2013; Fan et al., 2014) come into play to deal with noisy extractions. For this reason a recent trend of research has focused on Knowledge Base (KB) completion (Nickel et al., 2012; Bordes et al., 2013), exploiting the fact that distantly supervised OIE and structured knowledge can complement each other. However, the majority of integration approaches nowadays are not designed to deal with many different resources at the same time.

We propose an approach where the key idea is to bring together knowledge drawn from an arbitrary number of OIE systems, regardless of whether these systems provide links to some general-purpose inventory, come with their own ad-hoc structure, or have no structure at all. Knowledge from each source, in the form of ⟨subject, predicate, object⟩ triples, is disambiguated and linked to a single large sense inventory. This enables us to discover alignments at a semantic level between relations from different KBs, and to generate a unified, fully disambiguated KB of entities and semantic relations. KB-UNIFY achieves state-of-the-art disambiguation and provides a general, resource-independent representation of semantic relations, suitable for any kind of KB.

[1] http://lcl.uniroma1.it/kb-unify

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 726–736,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

The remainder of this paper is structured as follows: Section 2 reviews relevant related work; Sections 3, 4, 5 and 6 describe in detail each stage of the approach; Sections 7 and 8 describe the experiments carried out and the results obtained; and finally Section 9 summarizes our findings and discusses potential directions for future work.

2 Related Work

The integration of knowledge drawn from different sources has received much attention over the last decade. Among the most notable examples are resources like BabelNet (Navigli and Ponzetto, 2012), UBY (Gurevych et al., 2012) and YAGO (Mahdisoltani et al., 2015). While great effort has been put into aligning knowledge at the *concept level*, most approaches do not tackle the problem of integrating heterogeneous knowledge at the *relation level*, nor do they exploit effectively the huge amount of information harvested with OIE systems, even when this information is unambiguously linked to a structured resource, as in (Nakashole et al., 2012), or (Moro and Navigli, 2013). In fact, as the number of resources increases, KB alignment is already becoming an emergent research field: Dutta et al. (2014) describe a method for linking arguments in NELL triples to DBPEDIA by combining First Order Logic and Markov Networks; Grycner and Weikum (2014) semantify PATTY's pattern synsets and connect them to WordNet verbs; Lin et al. (2012) propose a method to propagate FREEBASE types across REVERB and deal with the problem of unlinkable entities. All these approaches achieve very competitive results in their respective settings, but unlike the approach being proposed here, they limit the task to 1-to-1 alignments. A few contributions have tried to broaden the scope and include different resources at the same time, but with rather different goals from ours. For example, Riedel et al. (2013) propose a universal schema that integrates structured data with OIE data by learning latent feature vectors for entities and relations; the KNOWLEDGE VAULT (Dong et al., 2014) uses a graph-based probabilistic framework where prior knowledge from existing resources (e.g. FREEBASE) improves Web extractions by predicting their reliability. However, in both cases the main objective is distantly supervised extraction from unstructured text, rather than KB unification. A recent trend of research focuses on learning embedding models for structured knowledge and their application to tasks like relation extraction and KB completion (Socher et al., 2013; Weston et al., 2013; Bordes et al., 2013). These approaches, however, leverage embeddings at surface level, which are suboptimal for our task, as will be discussed in Section 3. Since we require a common semantic framework for KB unification, we use vector representations based on word senses, which are mapped to a very large sense inventory. This shared sense inventory, then, constitutes the common ground in which disambiguation, alignment and final unification occurs.

3 Knowledge Base Unification: Overview

KB-UNIFY takes as input a set of KBs $\mathbf{K} = \{KB_1, ..., KB_n\}$ and outputs a single, unified and fully disambiguated KB, denoted as $\mathbf{KB}^*$. For our purposes we can define a KB KB_i as a triple $\langle E_i, R_i, T_i \rangle$, where E_i is a set of entities, R_i is a set of semantic relations, and T_i is a set of triples (facts) $\langle e_d, r, e_g \rangle$ with subject and object e_d, $e_g \in E_i$ and predicate $r \in R_i$. Depending on the nature of each KB_i, entities in E_i might be disambiguated and linked to an external inventory (e.g. the entity *Washington* linked to the Wikipedia page GEORGE WASHINGTON), or unlinked and only available as ambiguous mentions (e.g. the bare word *washington* might refer to the president, the city or the state). We can thus partition $\mathbf{K}$ into a subset of linked resources $\mathbf{K}_D$, and one of unlinked resources $\mathbf{K}_U$. In order to align very different and heterogeneous KBs at the semantic level, KB-UNIFY exploits:

- A unified sense inventory S, which acts as a superset for the inventories of individual KBs. We choose BabelNet (Navigli and Ponzetto, 2012) for this purpose: by merging complementary knowledge from different resources (e.g. Wikipedia, WordNet, Wikidata and Wiktionary, among others), BabelNet provides a wide coverage of entities and concepts whilst at the same time enabling convenient inter-resource mappings for KB_i in $\mathbf{K}_D$. For instance, each Wikipedia page (or Wikidata item) has a corresponding synset in BabelNet, which enables a one-to-one mapping between BabelNet's synsets and entries in, e.g., DBPEDIA or FREEBASE;

- A vector space model V_S that enables a semantic representation for every item in S. Current distributional models, like word em-

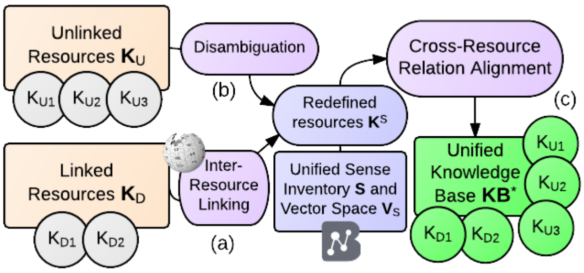


Figure 1: Unification algorithm workflow

beddings (Mikolov et al., 2013), are not suitable to our setting: they are constrained to surface word forms, and hence they inherently retain ambiguity of polysemous words and entity mentions. We thus leverage SENSEMBED (Iacobacci et al., 2015), a novel semantically-enhanced approach to embeddings. SENSEMBED is trained on a large annotated corpus and produces continuous representations for individual word senses (*sense embeddings*), according to an underlying sense inventory.

Figure 1 illustrates the workflow of our KB unification approach. Entities coming from any $KB_i \in \mathbf{K}_D$ can be directly (and unambiguously) mapped to the corresponding entries in S via BabelNet inter-resource linking (Figure 1(a)): in the above example, the entity *Washington* linked to the Wikipedia page GEORGE WASHINGTON is included in the BabelNet synset $\texttt{Washington}^4_{bn}$. In contrast, unlinked (and potentially ambiguous) entities need an explicit disambiguation step (Figure 1(b)) connecting them to appropriate entries, i.e. synsets, in S: this is the case, in the above example, for the ambiguous mention *washington* that has to be linked to either the president, the city or the state. Therefore, our approach comprises two successive stages:

- A **disambiguation** stage (Section 5) where all $KB_i \in \mathbf{K}$ are linked to S, either by inter-resource mapping (Figure 1(a)) or disambiguation (Figure 1(b)), and all E_i are merged into a unified set of entities E^*. As a result of this process we obtain a set $\mathbf{K}^S$ comprising all the KBs in $\mathbf{K}$ redefined using the common sense inventory S;

- An **alignment** stage (Section 6, Figure 1(c)) where, for each pair of KBs $KB^S_i, KB^S_j \in \mathbf{K}^S$, we compare any relation pair $\langle r_i, r_j \rangle$, $r_i \in R^S_i$ and $r_j \in R^S_j$, in order to identify cross-resource alignments and merge relations sharing equivalent semantics into relation clusters (*relation synsets*). This process yields a unified set of relation synsets R^*. The overall result is $\mathbf{KB}^* = \langle E^*, R^*, T^* \rangle$, where T^* is the set of all disambiguated triples redefined over E^* and R^*.

4 Background

The disambiguation stage of our approach is based on the interplay between two core components: a vector space model V_S, as introduced in Section 3, which provides an unambiguous semantic representation for each item in S; and a Word Sense Disambiguation/Entity Linking system, working on the same sense inventory S, which discovers and disambiguates concepts and entity mentions within a given input text. In this section we briefly describe our choice for these two components: SENSEMBED (Iacobacci et al., 2015) and BABELFY (Moro et al., 2014).

SENSEMBED is a knowledge-based approach for obtaining latent continuous representations of individual word senses. Unlike other sense-based embeddings approaches, like (Huang et al., 2012), which address the inherent polysemy of word-level representations relying solely on text corpora, SENSEMBED exploits the structured knowledge of a large sense inventory along with the distributional information gathered from text corpora. In order to do this, SENSEMBED requires a sense-annotated corpus; for each target word sense, then, a representation is computed by maximizing the log likelihood of the word sense with respect to its context within the annotated text, similarly to the word-based embeddings model. Following Iacobacci et al. (2015), we trained SENSEMBED using the English Wikipedia and, as sense inventory, BabelNet.

BABELFY[2] is a joint state-of-the-art approach to multilingual Entity Linking and Word Sense Disambiguation. Given the BabelNet lexicalized semantic network as underlying structure, BABELFY first models each concept in the network through its corresponding *semantic signature* by leveraging a graph random walk algorithm. Then, given an input text, the generated semantic signatures are used to construct a subgraph

[2] http://babelfy.org

of the semantic network representing the meaning of the content words in that text. BABELFY then searches this subgraph for the intended sense of each content word, by means of a densest-subgraph heuristic that identifies high-coherence interpretations. Given its unified approach that covers concepts and named entities alike, and its flexibility in disambiguating both bag-of-words and proper text, BABELFY constitutes the most convenient choice for linking relation triples to a high-coverage sense inventory like BabelNet.

5 Disambiguation

In the disambiguation phase (Figure 1(b)), all $KB_i \in \mathbf{K}_U$ are linked to the unified sense inventory S and added to the set of redefined KBs $\mathbf{K}^S$. As explained in Section 3, while each KB in $\mathbf{K}_D$ can be unambiguously redefined via BabelNet inter-resource links and added to $\mathbf{K}^S$, KBs in $\mathbf{K}_U$ require an explicit disambiguation step. Given $KB_i \in \mathbf{K}_U$, our disambiguation module (Figure 2) takes as input its set of unlinked triples T_i and outputs a set $T_i^S \subseteq T_i$ of disambiguated triples with subject-object pairs linked to S. The triples in T_i^S, together with their corresponding entity sets and relation sets, constitute the redefined KB_i^S which is then added to $\mathbf{K}^S$. However, applying a straightforward approach that disambiguates all triples in isolation might lead to very imprecise results, due to the lack of available context for each individual triple. We thus devised a disambiguation strategy that comprises three successive steps:

1. We identify a set of high-confidence seeds from T_i (Section 5.1), i.e. triples $\langle e_d, r, e_g \rangle$ where subject e_d and object e_g are highly semantically related, and disambiguate them using the senses that maximize their similarity in our vector space V_S;
2. We use the seeds to generate a ranking of the relations in R_i according to their degree of specificity (Section 5.2). We represent each $r \in R_i$ in our vector space V_S and assign higher specificity to relations whose arguments are closer in V_S;
3. We finally disambiguate the remaining non-seed triples in T_i (Section 5.3) starting from the most specific relations, and jointly using all participating argument pairs as context.

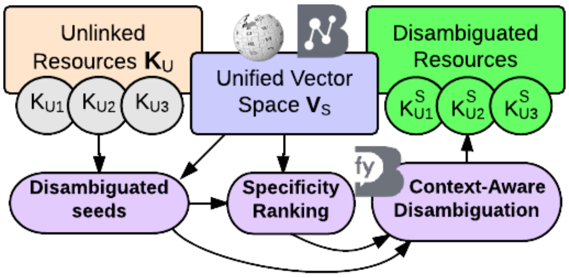


Figure 2: Disambiguation algorithm workflow

5.1 Identifying Seed Argument Pairs

The first stage of our disambiguation approach aims at extracting reliable seeds from T_i, i.e. triples $\langle e_d, r, e_g \rangle$ where subject e_d and object e_g can be confidently disambiguated without additional context. In order to do this we leverage the sense embeddings associated with each candidate disambiguation for e_d and e_g. We consider all the available senses for both e_d and e_g in S, namely $\mathbf{s}_d = \{s_d^1, ..., s_d^m\}$ and $\mathbf{s}_g = \{s_g^1, ..., s_g^{m'}\}$, and the corresponding sets of sense embeddings $\mathbf{v}_d = \{v_d^1, ..., v_d^m\}$ and $\mathbf{v}_g = \{v_g^1, ..., v_g^{m'}\}$. We then select, among all possible pairs of senses, the pair $\langle s_d^*, s_g^* \rangle$ that maximizes the cosine similarity between the corresponding embeddings $\langle v_d^*, v_g^* \rangle$:

$$\langle v_d^*, v_g^* \rangle = \mathrm{argmax}_{v_d \in \mathbf{v}_d,\, v_g \in \mathbf{v}_g} \frac{v_d \cdot v_g}{\|v_d\| \, \|v_g\|} \quad (1)$$

For each disambiguated triple $\langle s_d^*, r, s_g^* \rangle$, the cosine similarity value associated with $\langle v_d^*, v_g^* \rangle$ represents the disambiguation confidence ζ_{dis}. We rank all such triples according to their confidence, and select those above a given threshold δ_{dis}. The underlying assumption is that, for high-confidence subject-object pairs, the embeddings associated with the correct senses s_d^* and s_g^* will be closest in V_S compared to any other candidate pair. Intuitively, the more the relation r between e_d and e_g is semantically well defined, the more this assumption is justified. As an example, consider the triple ⟨*Armstrong*, *worked for*, *NASA*⟩: among all the possible senses for *Armstrong* (the astronaut, the jazz musician the cyclist, etc.) and *NASA* (the space agency, the racing organization, a Swedish band, etc.) we expect the vectors corresponding to the astronaut and the space agency to be closest in the vector space model V_S.

5.2 Relation Specificity Ranking

The assumption that, given an ambiguous subject-object pair, correct argument senses are the closest pair in the vector space (Section 5.1) is easily verifiable for general relations (e.g. *is a*, *is part of*). However, as a semantic relation becomes specific, its arguments are less guaranteed to be semantically related (e.g. *is a professor in the university of*) and a disambiguation approach based exclusively on similarity is prone to errors. On the other hand, specific relations tend to narrow down the scope of possible entity types occurring as subject and object. In the above example, *is a professor in the university of* requires entity pairs with professors as subjects and cities as objects. Our disambiguation strategy should thus vary according to the specificity of the relations taken into account. In order to consider this observation in our disambiguation pipeline, we first need to estimate the degree of specificity for each relation in the relation set R_i of the target KB to be disambiguated. Given R_i and a set of seeds from the previous stage (Section 5.1), we apply a specificity ranking policy and sort relations in R_i from the most general to the most specific. We compute the generality $Gen(r)$ of a given relation r by looking at the spatial dispersion of the sense embeddings associated with its seed subjects and objects. Let $\mathbf{v}_D$ ($\mathbf{v}_G$) be the set of sense embeddings associated to the domain (range) seed arguments of r. For both $\mathbf{v}_D$ and $\mathbf{v}_G$, we compute the corresponding centroid vectors μ_D and μ_G as:

$$\mu_k = \frac{1}{|\mathbf{v}_k|} \sum_{v \in \mathbf{v}_k} \frac{v}{\|v\|}, \quad k \in \{D, G\} \tag{2}$$

Then, the variances σ_D^2 and σ_G^2 are given by:

$$\sigma_k^2 = \frac{1}{|\mathbf{v}_k|} \sum_{v \in \mathbf{v}_k} (1 - cos\,(v, \mu_k))^2 \tag{3}$$

with $k \in \{D, G\}$ as before. We finally compute $Gen(r)$ as the average of σ_D^2 and σ_G^2. The result of this procedure is a *relation specificity ranking* that associates each relation r with its generality $Gen(r)$. Intuitively, we expect more general relations to show higher variance (hence higher $Gen(r)$), as their subjects and objects are likely to be rather disperse throughout the vector space; instead, arguments of very specific relations are more likely to be clustered together in compact regions, yielding lower values of $Gen(r)$.

5.3 Disambiguation with Relation Context

In the third step, both the specificity ranking and the seeds are exploited to disambiguate the remaining triples in T_i. To do this we leverage BABELFY (Moro et al., 2014) (introduced in Section 4). As we observed in Section 5.2, specific relations impose constraints on their subject-object types and tend to show compact domains and ranges in the vector space. Therefore, given a triple $\langle e_d, r, e_g \rangle$, knowing that r is specific enables us to put together all the triples in T_i where r occurs, and use them to provide meaningful context for disambiguation. If r is general, instead, its subject-object types are less constrained and additional triples do not guarantee to provide semantically related context.

At this stage, our algorithm takes as input the set of triples T_i, along with the associated disambiguation seeds (Section 5.1), the specificity ranking (Section 5.2) and a specificity threshold δ_{spec}. T_i is first partitioned into two subsets: T_i^{spec}, comprising all the triples for which $Gen(r) < \delta_{spec}$, and $T_i^{gen} = T_i \setminus T_i^{spec}$. We then employ two different disambiguation strategies:

- For each distinct relation r occurring in T_i^{spec}, we first retrieve the subset $T_{i,r}^{spec} \subset T_i^{spec}$ of triples where r occurs, and then disambiguate $T_{i,r}^{spec}$ as a whole with BABELFY. For each triple in $T_{i,r}^{spec}$, context is provided by all the remaining triples along with the disambiguated seeds extracted for r.
- We disambiguate the remaining triples in T_i^{gen} one by one in isolation with BABELFY, providing for each triple only the predicate string r as additional context.

6 Cross-Resource Relation Alignment

After disambiguation (Section 5) each KB in $\mathbf{K}$ is linked to the unified sense inventory S and added to $\mathbf{K}^S$. However, each $KB_i^S \in \mathbf{K}^S$ still provides its own relation set $R_i^S \subseteq R_i$. Instead, in the unified $\mathbf{KB}^*$, relations with equivalent semantics should be considered as part of a single relation synset even when they come from different KBs. Therefore, at this stage, we apply an alignment algorithm to identify pairs of relations from different KBs having equivalent semantics. We exploit the fact that each relation r is now defined over entity pairs linked to S, and we generate a semantic representation of r in the vector space V_S based on the centroid vectors of its domain and range. Due to representing the semantics of relations on this common ground, we can compare them by computing their domain and range similarity in V_S. We first consider each $KB_i^S \in \mathbf{K}^S$ and, for each relation r_i in R_i^S, we compute the corresponding centroid vectors $\mu_d^{r_i}$ and $\mu_g^{r_i}$ using formula (2). Then, for each pair of KBs $\langle KB_i^S, KB_j^S \rangle \in \mathbf{K}^S \times \mathbf{K}^S$, we compare all relation pairs $\langle r_i, r_j \rangle \in R_i^S \times R_j^S$ by computing the cosine similarity between domain centroids s_D and between range centroids s_G:

$$s_k = \frac{\mu_k^{r_i} \cdot \mu_k^{r_j}}{\|\mu_k^{r_i}\| \, \|\mu_k^{r_j}\|} \tag{4}$$

where μ_k^r denotes the centroid associated with relation r and $k \in \{D, G\}$. The average of s_D and s_G gives us an *alignment confidence* ζ_{align} for the pair $\langle r_i, r_j \rangle$. If confidence is above a given threshold δ_{align} then r_i and r_j are merged into the same relation synset. Relations for which no alignment is found are turned into singleton relation synsets. As a result of this alignment procedure we obtain the unified set of relations R^*.

7 Experimental Setup

The setting for our experimental evaluation was the following:

- We used BabelNet 3.0[3] as our unified sense inventory for the unification procedure as well as the underlying inventory for both BABELFY and SENSEMBED. Currently, BabelNet contains around 14M synsets and represents the largest single multilingual repository of entities and concepts;

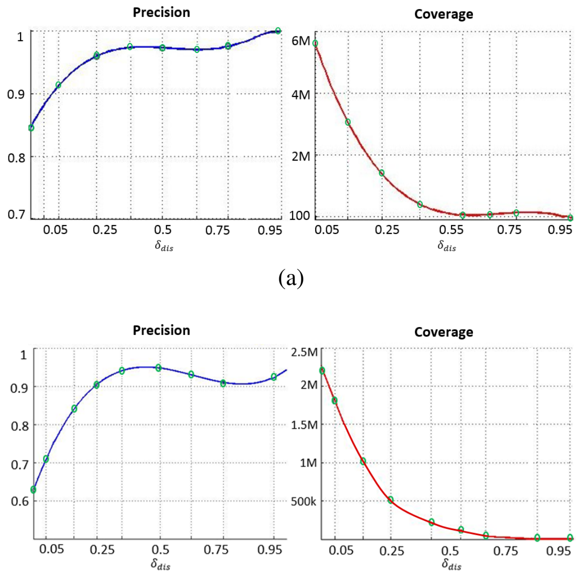
- We selected PATTY (Nakashole et al., 2012) and WISENET (Moro and Navigli, 2013) as linked resources. We used PATTY with FREEBASE types and pattern synsets derived from Wikipedia, and WISENET 2.0 with Wikipedia relational phrases;

- We selected NELL (Carlson et al., 2010) and REVERB (Fader et al., 2011) as unlinked resources. We used KB beliefs updated to November 2014 for the former, and the set of relation instances from ClueWeb09 for the latter.

Comparative statistics in Table 1 show that the input KBs are rather different in nature: NELL is based on 298 predefined relations and contains beliefs for about 2 million entities. The distribution of entities over relations is however very skewed, with 80.33% of the triples being instances of the `generalizations` relationship. In contrast, REVERB contains a highly sparse relation set (1,299,844 distinct relations) and more than 3 million distinct entities. PATTY features the largest (and, together with WISENET, sparsest) set of triples, with 1,631,531 distinct relations and less than 10 triples per relation on average.

	$\mathbf{K}_U$		$\mathbf{K}_D$	
	NELL	REVERB	PATTY	WISENET
# relations	298	1 299 844	1 631 531	245 935
# triples	2 245 050	14 728 268	15 802 946	2 271 807
# entities	1 996 021	3 327 425	1 087 907	1 636 307

Table 1: Statistics on the input KBs



(b)

Figure 3: Precision (left) and coverage (right) of disambiguated seeds at different values of δ_{dis} for (a) the whole set of triples in PATTY and (b) the subset of ambiguous triples

8 Experiments

8.1 Disambiguation

We tested our disambiguation approach experimentally in terms of both disambiguated seed quality (Section 8.1.1) and overall disambiguation performance (Section 8.1.2). We created a development set by extracting a subset of 6 million triples from the largest linked KB in our experimental setup, i.e. PATTY. Triples in PATTY are automatically linked to YAGO, which is in turn linked to WordNet and DBPEDIA. Since both resources are also linked by BabelNet, we mapped the original triples to the BabelNet sense inventory and used them to tune our disambiguation module. We also provide two baseline approaches: (1) di-

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

	SensEmbed			Baseline		
ζ_{dis}	0.5-0.7	0.7-0.9	0.9-1.0	0.5-0.7	0.7-0.9	0.9-1.0
Patty	.980	.980	1.000	.793	.780	1.000
WiSeNet	.958	.960	.973	.726	.786	.791
Nell	.955	.995	1.000	.800	.770	.885
ReVerb	.930	.940	.950	.775	.725	.920

Table 2: Disambiguation precision for all KBs

	$\delta_{spec} = 0.8$		$\delta_{spec} = 0.5$		$\delta_{spec} = 0.3$	
	all	only seeds	all	only seeds	all	only seeds
Patty	62.15	26.60	52.49	24.06	40.75	21.41
WiSeNet	60.00	37.46	54.44	22.26	53.58	16.62
Nell	76.97	62.98	50.95	20.71	44.70	4.36
ReVerb	41.20	38.57	25.14	23.70	13.37	12.75

Table 3: Coverage results (%) for all KBs

	KB-Unify		Dutta et al.	Baseline
	all	only seeds	($\alpha = 0.5$)	
Precision	.852	**.957**	.931	.749
Recall	**.875**	.117	.799	.608
F-score	**.864**	.197	.857	.671

Table 4: Disambiguation results over Nell gold standard

rect disambiguation on individual triples with Babelfy alone (without the seeds) and (2) direct disambiguation of the seeds only (without Babelfy).

8.1.1 Results: Disambiguated Seeds

We tuned our disambiguation algorithm by studying the quality of the disambiguated seeds (Section 5.1) extracted from the surface text triples of Patty. Figure 3 shows precision and coverage for increasing values of the confidence threshold δ_{dis}. We computed precision by checking each disambiguated seed against the corresponding linked triple in the development set, and coverage as the ratio of covered triples. We analyzed results for both the whole set of triples in Patty (Fig. 3a) and the subset of ambiguous triples (Fig. 3b), i.e. those triples whose subjects and objects have at least two candidate senses each in the BabelNet inventory. In both cases, precision of disambiguated seeds increases rapidly with δ_{dis}, stabilizing above 90% with $\delta_{dis} > 0.25$. Coverage displays the opposite behavior, decreasing exponentially with more confident outcomes, from 6 million triples to less than a thousand (for seeds with confidence $\delta_{dis} > 0.95$). As a result, we chose $\delta_{dis} = 0.25$ as optimal threshold value throughout the rest of the evaluations.

In addition, we manually evaluated the disambiguated seeds extracted from both linked KBs (Patty and WiSeNet) and unlinked KBs (Nell and ReVerb). For each KB, we extracted three random samples of 150 triples according to different levels of confidence ζ_{dis}: the first sample included extraction with $0.5 \leq \zeta_{dis} < 0.7$, the second with $0.7 \leq \zeta_{dis} < 0.9$, and the third with $\zeta_{dis} \geq 0.9$. Each sample was evaluated by two human judges: for each disambiguated triple $\langle e_d, r, e_g \rangle$, we presented our judges with the surface text arguments e_d, e_g and the relation string r, along with the two BabelNet synsets corresponding to the disambiguated arguments s^*_d, s^*_g, and we asked whether the association of each subject and object with the proposed BabelNet synset was correct. We then estimated precision as the average proportion of correctly disambiguated triples. For each sample we compared disambiguation precision using SensEmbed, as in Section 5.1, against the first baseline with Babelfy alone. Results, reported in Table 2, show that our approach consistently outperforms the baseline and provides high precision over all samples and KBs.

8.1.2 Results: Disambiguation with Relation Context

We then evaluated the overall disambiguation output after specificity ranking (Section 5.2) and disambiguation with relation context using Babelfy (Section 5.3). We analyzed three configurations of the disambiguation pipeline, namely $\delta_{spec} \in \{0.8, 0.5, 0.3\}$. We ran the algorithm over both linked and unlinked KBs of our experimental setup, and computed the coverage for each KB as the overall ratio of disambiguated triples. Results are reported in Table 3 and compared to the coverage obtained from the disambiguated seeds only: context-aware disambiguation substantially increases coverage over all KBs. Table 3 also shows that a restrictive δ_{spec} results in lower coverage values, due to the increased number of triples disambiguated without context.

Finally, we evaluated the quality of disambiguation on a publicly available dataset (Dutta et al., 2014) comprising manual annotations for Nell. This dataset provides a gold standard of 1200 triples whose subjects and objects are manually assigned a proper DBpedia URI. We again used BabelNet's inter-resource links to express DBpedia annotations with our sense inventory and then sought, for each annotated triple in the dataset, the corresponding triple in our disambiguated version of Nell with $\delta_{dis} = 0.25$ and $\delta_{spec} = 0.8$. We then repeated this process con-

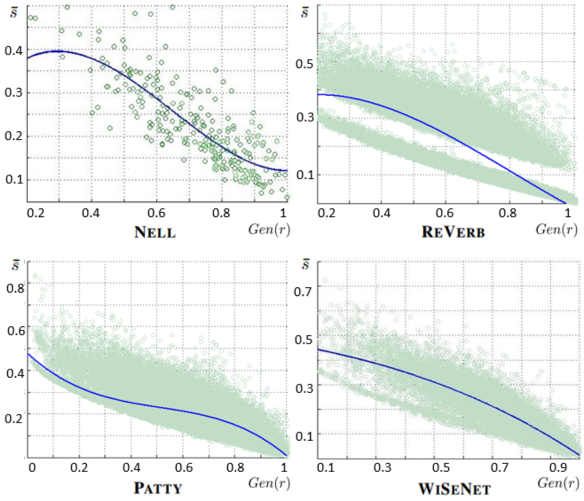


Figure 4: Average argument similarity against $Gen(r)$

	NELL	REVERB	PATTY	WISENET
Precision	.660	.715	.625	.750
Cohen's kappa	-	.430	.620	.600

Table 5: Specificity ranking evaluation

	NELL

High $Gen(r)$	`agent created`
	`at location`

Low $Gen(r)$	`person in economic sector`
	`restaurant in city`

	REVERB

High $Gen(r)$	`is for`
	`is in`

Low $Gen(r)$	`enter Taurus in`
	`carry oxygen to`

	PATTY

High $Gen(r)$	`located in`
	`later served to`

Low $Gen(r)$	`starting pitcher who played`
	`league coach for`

	WISENET

High $Gen(r)$	`include`
	`is a type of`

Low $Gen(r)$	`lobe-finned fish lived during`
	`took part in the Eurovision contest`

Table 6: Examples of general and specific relations for all KBs

sidering only the disambiguated seeds instead of the whole disambiguation pipeline. In line with (Dutta et al., 2014), we computed precision, recall and F-score for each setting. Results are reported in Table 4 and compared against those of Dutta et al. (2014) and against our first baseline with BABELFY alone. KB-UNIFY achieves the best result, showing that a baseline based on state-of-the-art disambiguation is negatively affected by the lack of context for each individual triple. In contrast, an approach that relies only on the disambiguated seeds affords very high precision, but suffers from dramatically lower coverage.

8.2 Specificity Ranking

We evaluated the specificity ranking (Section 5.2) generated by KB-UNIFY for all KBs of our experimental setup. First of all, we empirically validated our scoring function $Gen(r)$ over each resource: for each relation we computed the average cosine similarity among all its domain arguments $\bar{s}_D$ and among all its range arguments $\bar{s}_G$. We then plotted the average $\bar{s}$ of $\bar{s}_D$ and $\bar{s}_G$ against $Gen(r)$ for each relation r (Figure 4). As observed in Section 5.2, the average similarity among domain and range arguments decreases for increasing values of $Gen(r)$, indicating that more general relations allow less semantically constrained subject-object types.

We then used human judgement to assess the quality of our specificity rankings. First, each ranking was split into four quartiles, and two human evaluators were presented with a sample from the top quartile (i.e. a relation falling into the most general category) and a sample from the bottom quartile (i.e. a relation falling into the most specific category). We shuffled each relation pair, showed it to our human judges, and then asked which of the two relations they considered to be the more specific. Ranking precision was computed by considering those pairs where human choice agreed with the ranking. Finally, we computed inter-annotator agreement on each specificity ranking (except for NELL, due to the small sample size) with Cohen's kappa coefficient (Cohen, 1968). Results for each ranking are reported in Table 5, while some examples of general and specific relations for each KB are shown in Table 6. Disagreement between human choice and ranking is higher in NELL (where the set of relations is quite small compared to other KBs) and in PATTY (due to a sparser set of relations, biased towards very specific patterns). Inter-annotator agreement is instead lower for REVERB, where unconstrained Web harvesting often results in ambiguous relation strings.

8.3 Alignment

Due to the novelty of our approach, and hence the lack of widely accepted gold standards and testbeds, we evaluated our cross-resource relation alignment algorithm (Section 6) by exploiting hu-

	PATTY-WISENET		PATTY-REVERB		NELL-REVERB	
δ_{align}	0.7	0.9	0.7	0.9	0.7	0.9
Prec.	.68	**.80**	.58	**.74**	.61	**.75**
# Align.	128k	1.2k	47k	643	2.6k	88
	PATTY-NELL		WISENET-NELL		WISENET-REVERB	
δ_{align}	0.7	0.9	0.7	0.9	0.7	0.9
Prec.	.66	**1.00**	.70	**.84**	.59	**.87**
# Align.	2.6k	57	381	34	9.9k	169

Table 7: Cross-resource alignment evaluation

PATTY-WISENET		ζ_{align}
portrayed	's character	0.84
debuted in	first appeared in	0.86
PATTY-REVERB		ζ_{align}
language in	is spoken in	0.81
mostly known for	plays the role of	0.70
NELL-REVERB		ζ_{align}
academicfieldconcernssubject	is the scientific study of	0.75
drugpossiblytreatsphysiologicalcondition	is used to treat	0.92
NELL-PATTY		ζ_{align}
worksfor	was hired by	0.72
riveremptiesintoriver	tributary of	0.89
NELL-WISENET		ζ_{align}
animaleatfood	feeds on	0.72
teamhomestadium	play their home games at	0.88
REVERB-WISENET		ζ_{align}
has a selection of	offers	0.82
had grown up in	was born and raised in	0.85

Table 8: Examples of cross resource relation alignments

man judgement once again. Given the results of Section 8.1, we considered the top 10k frequent relations for each KB and ran the algorithm over each possible pair of KBs with two different configurations: $\delta_{align} = 0.7$ and $\delta_{align} = 0.9$. From each pair of KBs $\langle KB_i, KB_j \rangle$ we obtained a list of candidate alignments, i.e. pairs of relations $\langle r_i, r_j \rangle$ where $r_i \in KB_i$ and $r_j \in KB_j$. From each list we then extracted a random sample of 150 candidate alignments. We showed each alignment[4] $\langle r_i, r_j \rangle$ to two human judges, and asked whether r_i and r_j represented the same relation. The problem was presented in terms of paraphrasing: for each pair, we asked whether exchanging r_i and r_j within a sentence would have changed that sentence's meaning. In line with Section 8.2 we computed precision based on the agreement between human choice and automatic alignments. Results are reported in Table 7. Our alignment algorithm shows high precision in all pairings where $\delta_{align} = 0.9$. Alignment reliability decreases for lower δ_{align}, as relation pairs where r_i is a generalization of r_j (or vice versa) tend to have similar centroids in V_S. The same holds for pairs where r_i is the negation of r_j (or vice versa). Even though we could have utilized measures based on relation string similarity (Dutta et al., 2015) to reduce wrong alignments in these cases, by relying on a purely semantic criterion we removed any prior assumption on the format of input KBs. Some examples of alignments are shown in Table 8.

To conclude, we report statistics regarding the unified **KB*** produced from the initial set of resources in our experimental setup (cf. Section 7). We experimentally validated our thresholds for a high-precision setting, and then selected $\delta_{dis} = 0.25$, $\delta_{spec} = 0.8$ and $\delta_{align} = 0.8$. Our alignment algorithm produced 56,673 confident relation alignments, out of which 2,207 relation synsets were derived, with an average size of 16.82 individual relations per synset. As a result, we obtained a unified **KB*** comprising 24,221,856 disambiguated triples defined over 1,952,716 distinct entities and 2,675,296 distinct relations.

9 Conclusion and Future Work

We have presented KB-UNIFY, a novel, general approach for disambiguating and seamlessly unifying different KBs produced by different OIE systems. KB-UNIFY represents entities and relations using a shared semantic representation, leveraging a unified sense inventory together with a semantically-enhanced vector space model and a disambiguation algorithm. This enables us to disambiguate unlinked resources (like NELL and REVERB) with high precision and coverage, and to discover relation-level cross-resource alignments effectively. One of the key features of our strategy is its generality: by representing each KB on a common ground, we need no prior assumption on the nature and format of the knowledge it encodes. We tested our approach experimentally on a set of four very different KBs, both linked and unlinked, and we evaluated disambiguation and alignment results extensively at every stage, exploiting both human evaluations and public gold standard datasets (when available). This work opens compelling avenues for future work. We plan to further exploit sense-enhanced unified representations of relations in various ways: providing an ontological structure for the unified KB, exploring complementary approaches for capturing semantic relation alignments, and incorporating multilinguality.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

[4] In the case of relation synsets, such as PATTY and WISENET, we selected three random relation strings from each synset.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge. In *Proceedings of SIGMOD*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in NIPS*, volume 26, pages 2787–2795.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of AAAI*, pages 1306–1313.

Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220.

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 *Proceedings of the SIGKDD*, pages 601–610.

Arnab Dutta, Christian Meilicke, and Simone Paolo Ponzetto. 2014. A Probabilistic Approach for Integrating Heterogeneous Knowledge Sources. In *Proceedings of ESWC*, pages 286–301.

Arnab Dutta, Christian Meilicke, and Heiner Stuckenschmidt. 2015. Enriching Structured Knowledge with Open Information. In *Proceedings of WWW*, pages 267–277.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open Information Extraction from the Web. *Commun. ACM*, 51(12):68–74.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying Relations for Open Information Extraction. In *Proceedings of EMNLP*, pages 1535–1545.

Miao Fan, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, and Edward Y. Chang. 2014. Distant Supervision for Relation Extraction with Matrix Completion. In *Proceedings of ACL*, pages 839–849.

Adam Grycner and Gerhard Weikum. 2014. HARPY: Hypernyms and Alignment of Relational Paraphrases. In *Proceedings of ACL*, pages 2195–2204.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. Uby: A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of ACL*, pages 580–590.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of ACL*, pages 873–882.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proceedings of ACL*, pages 95–105.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, pages 1–29.

Thomas Lin, Mausam, and Oren Etzioni. 2012. No Noun Phrase Left Behind: Detecting and Typing Unlinkable Entities. In *Proceedings of EMNLP-CoNLL*, pages 893–903.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2015. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR*.

Tomas Mikolov, Kal Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of ACL-IJCNLP*, pages 1003–1011.

Andrea Moro and Roberto Navigli. 2013. Integrating Syntactic and Semantic Analysis into the Open Information Extraction Paradigm. In *Proceedings of IJCAI*, pages 2148–2154.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244.

Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek. 2012. PATTY: A Taxonomy of Relational Patterns with Semantic Types. In *Proceedings of EMNLP-CoNLL*, pages 1135–1145.

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.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2012. Factorizing YAGO: Scalable Machine Learning for Linked Data. In *Proceedings of WWW*, pages 271–280.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling Relations and Their Mentions Without Labeled Text. In *Proceedings of ECML-PKDD*, pages 148–163.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation Extraction with Matrix Factorization and Universal Schemas. In *Proceedings of NAACL*, pages 74–84.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning with Neural Tensor Networks for Knowledge Base Completion. In *Advances in NIPS*, pages 926–934.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance Multi-label Learning for Relation Extraction. In *Proceedings of EMNLP-CoNLL*, pages 455–465.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction. In *Proceedings of EMNLP*, pages 1366–1371.

Fei Wu and Daniel S. Weld. 2010. Open Information Extraction using Wikipedia. In *Proceedings of ACL*, pages 118–127.