Carnegie Mellon University

PIDGIN: Ontology Alignment using Web Text as Interlingua

Derry Wijaya, Partha Talukdar, Tom Mitchell

Language Technologies Institute, Carnegie Mellon University

Motivation, Goal and Contributions

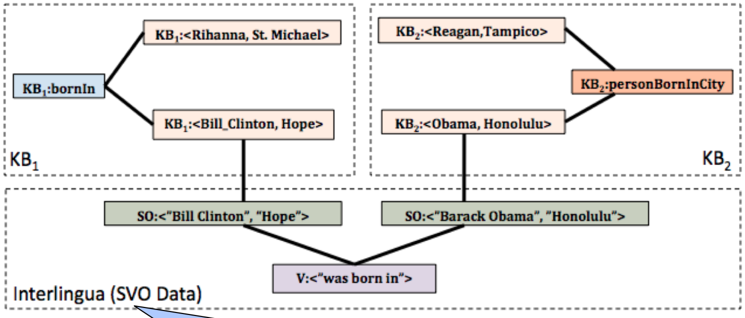
❑ **Motivation:** There are many knowledge sources independently developed, using different terminologies, coverage, and ontological structures in the semantic web: e.g., Wikipedia, Freebase, IMDB, NELL, Yago, etc

❑ **Goal:** There is a need to automatically align categories and relations across these knowledge bases (KBs): e.g., Yago:*wasBornIn* == NELL:*personbornincity*

❑ **Contributions:**

- ❑ **PIDGIN** - a novel graph-based ontology aligner that uses natural language text as an *interlingua* to align ontologies
- ❑ self-supervised and does not require human labeled data
- ❑ outperforms state-of-the-art ontology alignment system
- ❑ also learns verb phrases to identify relations and can type arguments of relations of different KBs

Motivating Example



600 million SVO triples from the entire ClueWeb corpus of about 230 billion tokens

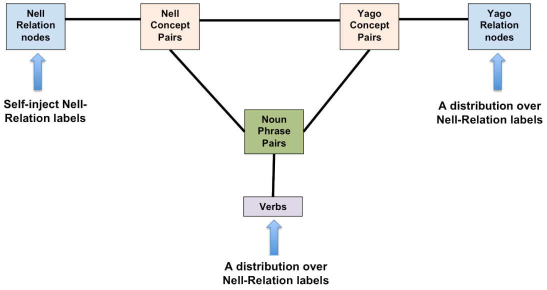
PIDGIN

❑ Analyzes both:

- ❑ the shared relation instances across KBs
- ❑ the mentions of relation instances across a large text corpus

❑ Treats alignment as classification over graph e.g., in the motivating example, **PIDGIN:**

- ❑ assigns (self) label to the node *KB$_1$:bornIn*
- ❑ starting with this initial label information, **PIDGIN** uses graph-based semi supervised learning to classify the rest of the nodes in the graph
- ❑ based on the assignment of scores of the labels on the KB$_2$ relation node, **PIDGIN** will determine the alignments between ontologies from these two KBs
- ❑ **PIDGIN** starts out by aligning relations from the two KBs, and produces category alignments as an important by product


Nell Relation nodes
Nell Concept Pairs
Yago Concept Pairs
Yago Relation nodes
Self-inject Nell-Relation labels
A distribution over Nell-Relation labels
Noun Phrase Pairs
Verbs
A distribution over Nell-Relation labels

Experiments

❑ We conduct experiments on several large scale, open domain, publicly available real world KBs:

- ❑ **NELL**: a large scale KB extracted automatically from web text
- ❑ **Yago2**: a large scale KB extracted automatically from semi-structured text of Wikipedia infoboxes
- ❑ **Freebase**: a large scale KB created collaboratively and manually by humans
- ❑ **KB Population (KBP)**: a smaller scale, manually constructed dataset used in the 2012 Text Analysis Conference for entity-linking, slot-filling and KB population tasks

❑ We compare Precision, Recall, and F1-score of resulting alignments with that of:

- ❑ **PARIS** (Probabilistic Alignment of Relations, Instances, and Schemas), a state-of-the-art ontology alignment system
- ❑ **JACCARD (inst)**, a baseline that computes the equivalence of relations using *Jaccard* similarity measures based on the number of overlap instances that the relations have

❑ We ask the questions:

- ❑ Whether **PIDGIN** improves precision, recall, and F1-score of relation and category alignments
- ❑ Whether adding more resources from text and more KBs as background knowledge in the graph improves alignment accuracy
- ❑ Whether **PIDGIN** is tolerant to noise
- ❑ What useful by-products of **PIDGIN** are
  - ❑ learn verb phrases to identify relations
  - ❑ type argument relations of different KBs
  - ❑ learn new relation instances

Results

Relation Alignments

KB Pair	System	Prec	Recall	F1
Freebase & NELL	JACCARD (inst)	0.61	0.51	0.56
	PARIS	0.47	0.09	0.15
	PIDGIN	**0.65**	**0.61**	**0.63**
Yago2 & NELL	JACCARD (inst)	0.56	0.43	0.49
	PARIS	**0.67**	0.09	0.15
	PIDGIN	0.52	**0.52**	**0.52**
KBP & NELL	JACCARD (inst)	0.0	0.0	0.0
	PARIS	0.0	0.0	0.0
	PIDGIN	**0.07**	**0.06**	**0.06**

❑ Examples of relation alignments: Freebase: */medicine/medical_treatment/side_effects* == NELL:*drughassideeffect*, Freebase:*/sports/league/arena_stadium* == NELL:*leaguestadiums*

❑ Examples of category alignments: Yago2:*wordnet_actor_109765278* == NELL:*actor*, Yago2:*yagoURL* == NELL:*website*, Yago2:*yagoLegalActor* == NELL:*agent*

Learned Verbs that Identify Relations

Knowledge Base	Relation	Learned Verbs
Freebase	/sports/sports_team/arena_stadium	played at, played in, defeated at, will host at, beaten at
	/medicine/medical_treatment/side_effects	may promote, can cause, may produce, is worsen, exacerbate
NELL	drugPossiblyTreats PhysiologicalCondition	treat, relieve, reduce, help with, can help alleviate
	politicianHoldsOffice	serves as, run for, became, was elected
Yago2	actedIn	played in, starred in, starred, played, portrayed in
	isMarriedTo	married, met, date, wed, divorce

Conclusion and Future Works

❑ **PIDGIN** improves recall using text-based interlingua to establish alignments when there are few or no shared relation instances between the KBs

❑ **PIDGIN** automatically learns verbs to identify relations that can be used to extract new instances to populate the KB or identify relations between two entities in a document

❑ **PIDGIN** automatically types arguments of relations e.g., Freebase:*/business/industry/name* is typed with NELL <*company*, *economicsector*>. Argument typing can improve the accuracy of extraction of new relation instances by constraining the instances to have the correct types