Wordnet Creation and Extension Made Simple

A multilingual lexicon-based approach using Wiki resources

Valérie Hanoka [1,2] Benoît Sagot [1]

1. Alpage, INRIA Paris-Rocquencourt & Université Paris 7
2. Verbatim Analysis

LREC'12 – Istanbul
May 25, 2012

Wordnet creation and extension

- **Lexical semantic knowledge** in NLP research and applications are commonly represented using **wordnets**
- Developing wordnets is **time-consuming** and **expensive**
- Need for techniques that use **easily extractible multilingual lexical resources**

Related Work

Publication	Parallel Corpora	Translations Lexicons	Glosses Context	Back -trans.	Machine Learning	Wiki Resources	Bi-lingual	Multi-lingual
Dyvik (1998)	✓			✓			✓	
Pianta et al. (2002)		✓	✓	✓			✓	
Sagot and Fišer (2008)	✓	✓				✓	✓	✓
De Melo and Weikum (2009)	✓	✓	✓		✓	✓		✓
Navigli and Ponzetto (2010)			✓			✓		✓
This work				✓		✓		✓

This Work

General aim

A technique to **create** or **improve** new wordnets in many languages based on a **large-scale multilingual translation/synonymy graph** extracted from **Wiktionaries** and **Wikipedia**

Experiments on French

- **Bootstrapping** of a new French wordnet
- **Comparison** with the **Wolf**,
  a Free French Wordnet (Sagot and Fišer, 2008)
- The results were used to **extend the Wolf**.

Methodology

Process Outline

Inputs

- Large-scale multilingual Translation/Synonymy directed graph involving many languages
- Synset-aligned wordnets in m different languages ideally $m \geq 3$

Translation scoring

Output

Synset-aligned wordnets with ranked translations:
m extended wordnets + additional new wordnets for other languages

Building a large-scale multilingual translation graph

Translation and synonym pairs extraction

- **Wiktionaries** in 18 languages:
  Czech, Dutch, English, French, German, Hebrew, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Romanian, Russian, Slovak, Spanish, Swedish and Turkish
- **French Wikipedia**

9.91M translations pairs and **0.75M** synonymy pairs

Building a large-scale multilingual translation graph

Filtering

- Removing **duplicates**
- Removing **under-represented terms**
  terms that are not present in at least n translations (i.e $n = 3$)

1.58M translations pairs and **0.37M** synonymy pairs

Average degree of the multilingual translation graph (both directions): **3.52**

Filling Synsets

We redefine the notions of literals and synsets in a multilingual setting involving aligned wordnets.

- Let a **literal** be a triplet (**term**, **language**, **weight**)
- Let a **synset** be the union of all literals present in the synsets of all input wordnets that correspond to a same ID

English	Romanian	Bulgarian
inequality	imparitate inegalitate	неравенство

{(inequality, en, W_{max}), (imparitate, ro, W_{max}), (inegalitate, ro, W_{max}), (неравенство, bg, W_{max})}

Filling Synsets

Generating Candidates

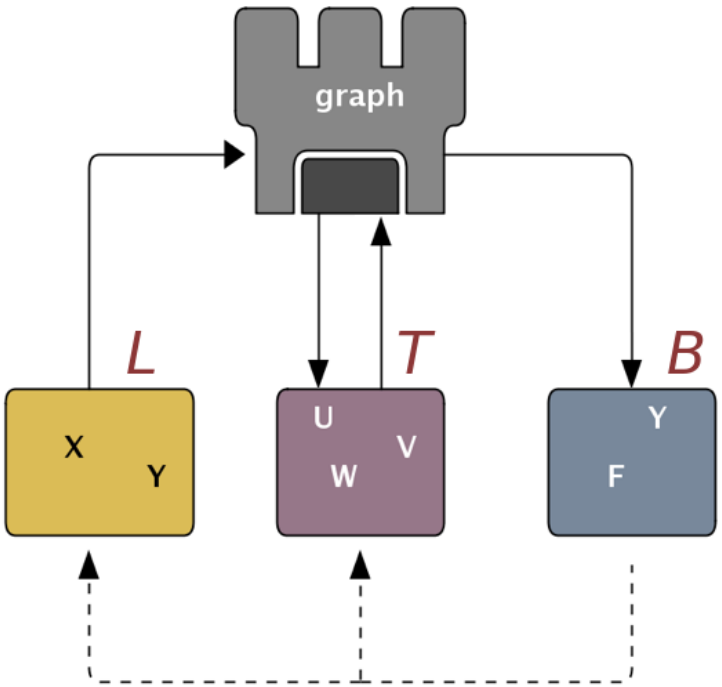
Let L_0 be an aligned synset of **gold literals**

{(inequality, en, W_{max}), (imparitate, ro, W_{max}), (inegalitate, ro, W_{max}), (неравенство, bg, W_{max}) }

L_0 is used to query the translation graph in order to propose a new multilingual set T of **candidate literals**.

Candidate literals' weights are updated according to the quality of their back-translation.

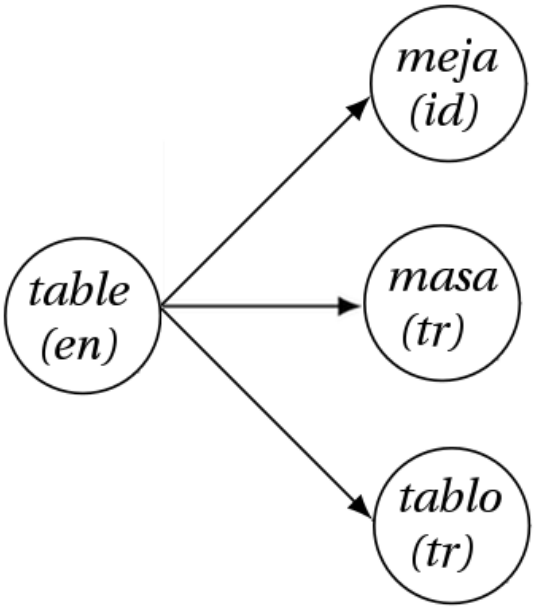
Back-Translation Algorithm



for each litteral $\mathfrak{l} \in L_i$:

1. $T \leftarrow$ getTranslationsFor($\mathfrak{l}$)
   for each translation $\mathfrak{t} \in T$:
   - $B \leftarrow$ getTranslationsFor($\mathfrak{t}$)
   - Update $\mathfrak{t}$'s weight according to the score of the translation $(\mathfrak{l}, \mathfrak{t})$ in the graph.
2. for each back-translation $\mathfrak{b} \in B$:
   - Whether $\mathfrak{b} \in L_i$: (in|de)crement heuristically the weight of $\mathfrak{t}$
3. $L_{i+1} \leftarrow L_i \cup T$

Back-Translation Algorithm



$deg_{tr}(\text{table}) = 2$
$deg_{id}(\text{table}) = 1$

How translations are scored in the graph ?

Translation from (A, $lang_1$) to (C, $lang_2$)
is scored $\frac{1}{deg_{lang_2}(A)}$.

Experiments

Input

- Synset-aligned wordnets in 4 languages: Bulgarian, Czech, English and Romanian + an empty wordnet in French
- Multilingual Translation/Synonymy directed graph containing 1.58M translations pairs and 0.37M synonymy pairs extracted from 18 wiktionaries

Output

Synset-aligned extended wordnets with scored translations

English	Romanian	Bulgarian	French
inequality triangle inequality(2.0) dissimilarity(1.5)	**imparitate** **inegalitate**	**неравенство**	inégalité (12.0) dissemblance (1.0) inéquation (1.0)

Results: a few examples

French literal	Score	PWN literals in the synset	PWN definition	Correct	Already in WOLF
ensemble	66.3	together	in each other's company	✓	×
permettre	77.0	allow, permit, tolerate	allow the presence of or allow without opposing or prohibiting	✓	×
à peu près	72.9	about, just about, almost, most, all but, nearly,[...]	slightly short of or not quite accomplished	✓	×
ivre	100.0	intoxicated, drunk	as if under the influence of alcohol	✓	×
périlleux	40.6	hazardous, risky, venturesome, venturous	involving risk or danger	✓	✓
accord	34.7	agreement	the verbal act of agreeing	✓	✓
tête	100.0	head	the top of something	✓	✓
tête	46.1	drumhead, head	a membrane that is stretched taut over a drum	×	×
salamandre	35.6	poker, stove poker, fire, hook, salamander	fire iron consisting of a metal rod with a handle	×	×

Evaluation

Evaluation and Comparison with the Wolf

Two parameters for the evaluation:

t $\rightarrow$ threshold on the score

n_{max} $\rightarrow$ upper bound (keeps the n best candidates for each synset)

We considered only (literal, synset) candidates with $t > 30$
(10,568 candidates, among which 58% are not already in the WOLF)

Manual evaluation
of 400 randomly chosen (literal, synset) candidates

Evaluation

$$Precision = \frac{|\{\text{correct candidates}\}|}{|\{\text{all candidates retained}\}|}$$

	$t = 30$	$t = 40$	$t = 50$	$t = 60$
$n_{\max} = 1$	8362/**77.3**	5340/**81.5**	3353/**85.6**	2245/**90.5**
$n_{\max} = 3$	10403/**74.8**	6298/**80.6**	3890/**85.1**	2582/**89.6**
$n_{\max} = \infty$	10568/**74.1**	6357/**80.3**	3917/**85.2**	2594/**89.6**

Estimation of the number of candidates / **Precision**

Conclusion and perspectives

- Generated candidates by exploiting a large **highly multilingual** translation graph extracted from a set of **wiki resources**
- High or medium frequency words that are **polysemous** — Not generated by previous approaches.
  'permettre/allow', 'manger/to eat', 'taper/to hit', 'lent/slow'
- Well suited for creating or enriching wordnets for languages that have at their disposal **large or medium coverage** Wiktionary and Wikipedia

- Perspectives:
  1. upgrading the translation graph's quality as in Mausam et al. (2009)
  2. computing scores for candidate literals by doing a **walk** in the translation graph.

Thank You