Using Wikipedia for Named-Entity Translation

Izaskun Fernandez, Iñaki Alegria, Nerea Ezeiza

Tekniker-IK4 and IXA NLP Group

Index

1. Introduction and Motivation
2. Related Works and Previous Works
3. System Development
4. Results
5. Conclusion and future work

Index

1. Introduction and Motivation
2. Related Works and Previous Works
3. System Development
4. Results
5. Conclusion and future work

Introduction and Motivation I

- Named Entities: person, location, organization
- Named Entities (NE) Recognition: common task
- Main Goal: Construct a multilingual NE database
  - translation systems
  - multilingual information extraction (QA)
- NE translations

Introduction and Motivation II

Exploit Wikipedia for NE translation

- Free on-line multilingual encyclopedia
- Each entry uniquely represented by its title
- Wikipedia Interlingual Links (WIL) to relate same titles in different languages:
    - Basque: *Euskal Herria*
    - English: *Basque Country*

Index

1. Introduction and Motivation
2. Related Works and Previous Works
3. System Development
4. Results
5. Conclusion and future work

Related Works I

NE translation:

- English-French translation system based on parallel corpora using statistical methods: *Learning Translations of Named-Entity Phrases from Parallel Corpora (Moore R., EACL 2003)*
- Arabic-English translation system based on comparable corpora using simple transformation rules and dictionaries: *Machine Transliteration of Names in Arabic Text (Al-Onaizan Y. et al., ACL 2002)*

Related Works II

Wikipedia and NE:

- Classification based approach for German-English WILs enrichment: *Enriching the Crosslingual Link Structure of Wikipedia (Sorg P., and Cimiano T. AAAI2008)*
- Multilingual NER based on Wikipedia, exploiting English data for bootstrap NER process in other languages: *Mining Wiki Resources for Multilingual Named Entity Recognition (Richman A. E., Schone P. ACL2008)*

Previous Work I

Basque-Spanish language dependent system

- Basque-Spanish comparable corpora
- Linguistic knowledge based transliteration module (phonetic/phonological information)
- Linguistic knowledge based re-arrangement module (morphosyntactic information)

Previous Work II

Language semi-independent translation tool

- Basque-Spanish
- Spanish-English

Using:

- Comparable corpora for each language pair
- Bilingual dictionary for each language pair
- Edit distance based transliteration module
- Re-arrangement module: all with all

Index

1. Introduction and Motivation
2. Related Works and Previous Works
3. System Development
4. Results
5. Conclusion and future work

Resources

Resources for constructing translation tool:

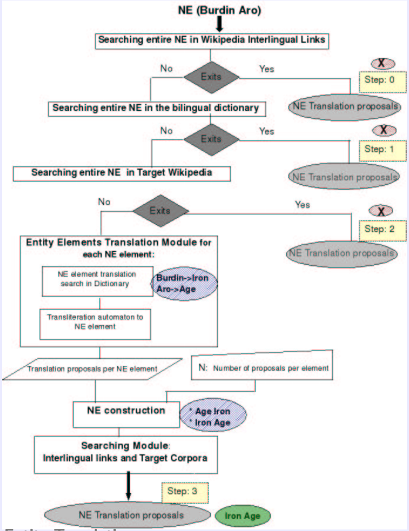
- MediaWiki API (`http://www.mediawiki.org/wiki/API`): WIL and redirection pages
- Yahoo! semantically annotated Wikipedia version: target lexicon using only NEs (`http://www.yr-bcnn.es/semanticWikipedia`)
- Basque-English bilingual dictionary

System Description

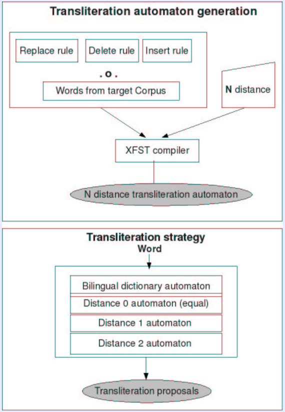
Three main modules:

- Searching module:
    - wikipedia interlingual links
    - bilingual dictionary
    - target NEs from Wikipedia
- Entity element translation module:
    - transliteration grammar
    - bilingual dictionary
- Element arranging module
    - all with all combinations

System Architecture


NE (Burdin Aro)
Searching entire NE in Wikipedia Interlingual Links
No
Exits
Yes
X
Step: 0
NE Translation proposals
Searching entire NE in the bilingual dictionary
No
Exits
Yes
X
Step: 1
NE Translation proposals
Searching entire NE in Target Wikipedia
No
Exits
Yes
X
Step: 2
NE Translation proposals
Entity Elements Translation Module for each NE element:
NE element translation search in Dictionary
Burdin->Iron
Aro->Age
Transliteration automaton to NE element
Translation proposals per NE element
N: Number of proposals per element
NE construction
* Age Iron
* Iron Age
Searching Module: Interlingual links and Target Corpora
Step: 3
NE Translation proposals
Iron Age

Entity Elements Translation Strategy


Transliteration automaton generation
Replace rule
Delete rule
Insert rule
. o .
Words from target Corpus
N distance
XFST compiler
N distance transliteration automaton
Transliteration strategy
Word
Bilingual dictionary automaton
Distance 0 automaton (equal)
Distance 1 automaton
Distance 2 automaton
Transliteration proposals

Index

1. Introduction and Motivation
2. Related Works and Previous Works
3. System Development
4. Results
5. Conclusion and future work

Experimental Settings

- Two evaluation corpus:
  - Most frequent NEs at *Egunkaria 2002* data-set with Basque-English WILs
    - 142,464 NEs in data-set
    - 575 NE filtered for evaluation
    - WILs for tranlation (Step0) not used for this evaluation
  - ResPubliQA CLEF2009 test set
    - 500 questions: Bulgarian, Basque, English, etc.
    - 72 Basque-English NE pairs
    - 9 of them without entry in the target Wikipedia

Measures

Three measures:

- $Precision = \frac{correctly\_translated\_NEs}{Translated\_NEs}$
- $Recall = \frac{correctly\_translated\_NEs}{All\_NEs}$
- $F - score = \frac{2*Precision*Recall}{Precision+Recall}$

Baseline:

- correct translation when Basque and English forms are identical

Evaluation with Wikipedia based test set I

Translation distribution:

Steps	Total	Correct
Bilingual dictionary	17	11
Target Wikipedia	391	375
Element by element	59	48
No Translation	108	0

Results:

	Pr	R	fs
Baseline	59.82%	59.82%	59.82%
Our system	93.36%	75.82%	83.68%

Evaluation with Wikipedia based test set II

Encouraging results: 83.68% f-score

Analysing errors: WILs not always link equivalent forms

- WIL: *Dorre bikiak - World Trade Center*
- Correct link: *Dorre bikiak - Twin Towers*
- New WIL suggestion

Evaluation with CLEF2009 based test set I

9 of the 72 pairs without entry in the target Wikipedia

Topline recall 87.5%

Evalution in two different ways:

- silence-mode: when no proposal found, no translation is returned
- talkative-mode: when no proposal found, the original Basque form is proposed

Results:

	Pr	**R**	**fs**
Baseline	23.69%	23.69%	23.69%
Silence-mode	92.68%	52.77%	67.25%
Talkative-mode	55.5%	55.5%	55.5%

Evaluation with CLEF2009 based test set II

CLEF2009 set not belong to Wikipedia

- WILs explotation for NE translation (Step0)
- 26 NE translation proposal

System improvement respect to the baseline

System Improvement I

Not very suitable bilingual dictionary

- *Nazio Batuak - United Nations*
- Using the dictionary: Union and Nation

Automatic dictionary lexical enrichment:

- WILs of 84 wrong translated NE pairs in the Wikipedia-based test set
- For each NE pair: try to match each element Basque form with their English form:
  - mantaining the original Basque form
  - using the existing bilingual dictionary
  - when every Basque element except one parsed, and only one English element has no Basque element assigned, enrich dictionary with the new pair
- Iterate until no new word pair is found

System Improvement II

Example:*Europako Parlamentua - European Parliament*

- Try *Europako*. No equivalence found
- Try *Parlamentua*. Bilingual dictionary: *Parliament*
- Without matching *Europako* and *European*. Add new pair to dictionary

Results:

	Pr	**R**	**fs**
Silence-mode	92.68%	52.77%	67.25%
Talkative-mode	55.5%	55.5%	55.5%
Enriched-Silence-mode	**93%**	**55.5%**	**69.51%**
Enriched-Talkative-mode	**58.33%**	**58.33%**	**58.33%**

Index

1. Introduction and Motivation
2. Related Works and Previous Works
3. System Development
4. Results
5. Conclusion and future work

Conclusions and Future Works

- Exploiting Wikipedia for NE translation might benefit in two directions:
  - Building a good-quality NE translation system
  - Suggesting new WILs
- Promising results but deeper evaluation and error analysis is needed
- Future works:
  - NE disambiguation, specially for minority languages
  - Using the presented NE translation system for that purpose

Using Wikipedia for Named-Entity Translation

Izaskun Fernandez, Iñaki Alegria, Nerea Ezeiza

Tekniker-IK4 and IXA NLP Group