Automatic acquisition of Named Entities for Rule-Based Machine Translation

Antonio Toral, Andy Way – DCU

2^{nd} International Workshop on Free/Open-Source Rule-Based Machine Translation

2011/01/20

Contents

- Named Entities (NEs) refer to proper nouns (e.g. person, location, organization). Information Extraction, MUC

- Named Entities (NEs) refer to proper nouns (e.g. person, location, organization). Information Extraction, MUC
- Distribution of NEs compared to other PoS

- Named Entities (NEs) refer to proper nouns (e.g. person, location, organization). Information Extraction, MUC
- Distribution of NEs compared to other PoS
  - English Europarl, tagged with Freeling

- Named Entities (NEs) refer to proper nouns (e.g. person, location, organization). Information Extraction, MUC
- Distribution of NEs compared to other PoS
  - English Europarl, tagged with Freeling
  - Mean: 1 NEs, 3 common nouns, 7 verbs

- Named Entities (NEs) refer to proper nouns (e.g. person, location, organization). Information Extraction, MUC
- Distribution of NEs compared to other PoS
  - English Europarl, tagged with Freeling
  - Mean: 1 NEs, 3 common nouns, 7 verbs
  - Avg num occurrences: 24 NEs, 295 common nouns, 888 verbs

- Named Entities (NEs) refer to proper nouns (e.g. person, location, organization). Information Extraction, MUC
- Distribution of NEs compared to other PoS
  - English Europarl, tagged with Freeling
  - Mean: 1 NEs, 3 common nouns, 7 verbs
  - Avg num occurrences: 24 NEs, 295 common nouns, 888 verbs
  - Num different instances: 88k NEs, 26k common nouns, 7k verbs

- Multilingual and Interoperable Named Entity Lexicon (MINELex)

- Multilingual and Interoperable Named Entity Lexicon (MINELex)
- NEs acquired from Wikipedia for 11 languages and connected to LRs:

- Multilingual and Interoperable Named Entity Lexicon (MINELex)
- NEs acquired from Wikipedia for 11 languages and connected to LRs:
  - Semantic units of dictionaries (en, es, it, ar). E.g. Tim_Robbins instance-of actor, film_director

- Multilingual and Interoperable Named Entity Lexicon (MINELex)
- NEs acquired from Wikipedia for 11 languages and connected to LRs:
  - Semantic units of dictionaries (en, es, it, ar). E.g. Tim_Robbins instance-of actor, film_director
  - Nodes of ontologies (SUMO, SIMPLE). E.g. Tim_Robbins + Position, believes

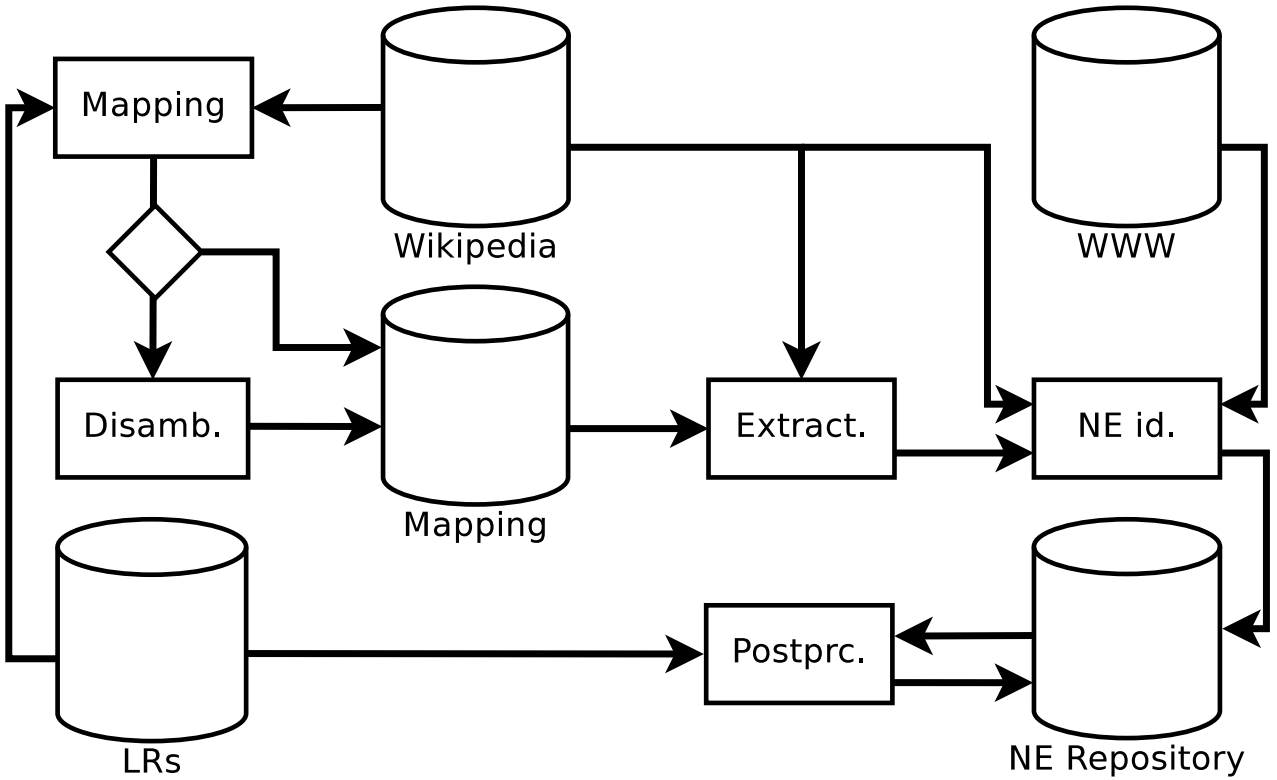
- Multilingual and Interoperable Named Entity Lexicon (MINELex)
- NEs acquired from Wikipedia for 11 languages and connected to LRs:
  - Semantic units of dictionaries (en, es, it, ar). E.g. Tim_Robbins instance-of actor, film_director
  - Nodes of ontologies (SUMO, SIMPLE). E.g. Tim_Robbins + Position, believes
- Equivalent NEs in different languages connected by interlingual links

- Multilingual and Interoperable Named Entity Lexicon (MINELex)
- NEs acquired from Wikipedia for 11 languages and connected to LRs:
  - Semantic units of dictionaries (en, es, it, ar). E.g. Tim_Robbins instance-of actor, film_director
  - Nodes of ontologies (SUMO, SIMPLE). E.g. Tim_Robbins + Position, believes
- Equivalent NEs in different languages connected by interlingual links
- NEs associated with confidence scores (num occurrences, % occurs capitalised)

- Multilingual and Interoperable Named Entity Lexicon (MINELex)
- NEs acquired from Wikipedia for 11 languages and connected to LRs:
  - Semantic units of dictionaries (en, es, it, ar). E.g. Tim_Robbins instance-of actor, film_director
  - Nodes of ontologies (SUMO, SIMPLE). E.g. Tim_Robbins + Position, believes
- Equivalent NEs in different languages connected by interlingual links
- NEs associated with confidence scores (num occurrences, % occurs capitalised)

- Multilingual and Interoperable Named Entity Lexicon (MINELex)
- NEs acquired from Wikipedia for 11 languages and connected to LRs:
  - Semantic units of dictionaries (en, es, it, ar). E.g. Tim_Robbins instance-of actor, film_director
  - Nodes of ontologies (SUMO, SIMPLE). E.g. Tim_Robbins + Position, believes
- Equivalent NEs in different languages connected by interlingual links
- NEs associated with confidence scores (num occurrences, % occurs capitalised)

	English	Spanish
NEs	948,410	99,330
Variants	1,541,993	128,796
Instance relations	1,366,899	128,796


Mapping
Wikipedia
WWW
Disamb.
Mapping
Extract.
NE id.
LRs
Postprc.
NE Repository

- Aim: automatically add NEs to RBMT dictionaries

- Aim: automatically add NEs to RBMT dictionaries
- Reasons:

- Aim: automatically add NEs to RBMT dictionaries
- Reasons:
  - Distributional properties + dynamic nature of NEs $\rightarrow$ impractical to build dictionaries manually

- Aim: automatically add NEs to RBMT dictionaries
- Reasons:
  - Distributional properties + dynamic nature of NEs → impractical to build dictionaries manually
  - 1/3 of entries in Apertium English–Spanish dic regard NEs

- Aim: automatically add NEs to RBMT dictionaries
- Reasons:
  - Distributional properties + dynamic nature of NEs → impractical to build dictionaries manually
  - 1/3 of entries in Apertium English–Spanish dic regard NEs
  - Simpler morphology of NEs*

- Aim: automatically add NEs to RBMT dictionaries
- Reasons:
  - Distributional properties + dynamic nature of NEs → impractical to build dictionaries manually
  - 1/3 of entries in Apertium English–Spanish dic regard NEs
  - Simpler morphology of NEs*

- Extract bilingual pairs of NEs from MINELex and insert into Apertium dics

- Extract bilingual pairs of NEs from MINELex and insert into Apertium dics
- Subset of NEs that satisfy restrictions

- Extract bilingual pairs of NEs from MINELex and insert into Apertium dics
- Subset of NEs that satisfy restrictions
  - Min num of occurrences

- Extract bilingual pairs of NEs from MINELex and insert into Apertium dics
- Subset of NEs that satisfy restrictions
  - Min num of occurrences
  - Min % of occurrences are capitalised

- Extract bilingual pairs of NEs from MINELex and insert into Apertium dics
- Subset of NEs that satisfy restrictions
  - Min num of occurrences
  - Min % of occurrences are capitalised
- Insert relevant data in Apertium dics (sl, tl, bi)

NE English = Yekaterinburg
NE Spanish = Ekaterimburgo
Number occurrences = 190
Percentage capitalised = .95

```
<pardef n="Aachen_np">
 <e><p><l/><r>
      <s n="np"/><s n="al"/><s n="sp"/>
  </r></p></e>
</pardef>
<e lm="Yekaterinburg">
 <i>Yekaterinburg</i>
 <par n="Aachen__np"/>
</e>
```

```
NE English = Yekaterinburg
NE Spanish = Ekaterimburgo
Number occurrences = 190
Percentage capitalised = .95
```

NE English = Yekaterinburg
NE Spanish = Ekaterimburgo
Number occurrences = 190
Percentage capitalised = .95

```
<pardef n="Aachen_np">
 <e><p><l/><r>
      <s n="np"/><s n="al"/><s n="sp"/>
  </r></p></e>
</pardef>
<e lm="Yekaterinburg">
 <i>Yekaterinburg</i>
 <par n="Aachen__np"/>
</e>
```

```
<pardef n="Aquisgran_np">
 <e><p><l/><r>
  <s n="np"/><s n="al"/><s n="mf"/><s n="sp"/>
 </r></p></e>
</pardef>
<e lm="Ekaterimburgo">
 <i>Ekaterimburgo</i>
 <par n="Aquisgran__np"/>
</e>
```

NE English = Yekaterinburg
NE Spanish = Ekaterimburgo
Number occurrences = 190
Percentage capitalised = .95

```
<pardef n="Aachen_np">
 <e><p><l/><r>
      <s n="np"/><s n="al"/><s n="sp"/>
  </r></p></e>
</pardef>
<e lm="Yekaterinburg">
 <i>Yekaterinburg</i>
 <par n="Aachen__np"/>
</e>
```

```
<pardef n="Aquisgran_np">
 <e><p><l/><r>
  <s n="np"/><s n="al"/><s n="mf"/><s n="sp"/>
 </r></p></e>
</pardef>
<e lm="Ekaterimburgo">
 <i>Ekaterimburgo</i>
 <par n="Aquisgran__np"/>
</e>
```

```
<e><p>
  <l>Yekaterinburg
   <s n="np"/><s n="al"/>
  </l>
  <r>Ekaterimburgo
   <s n="np"/><s n="al"/><s n="mf"/>
  </r>
</p></e>
```

- System: Apertium es–en 0.7.1

- System: Apertium es–en 0.7.1
- Baselines: Apertium without NEs (no_nes) and Apertium with handtagged NEs (nes)

- System: Apertium es–en 0.7.1
- Baselines: Apertium without NEs (no_nes) and Apertium with handtagged NEs (nes)
- Test set: nc-2007 (WMT'08)

- System: Apertium es–en 0.7.1
- Baselines: Apertium without NEs (no_nes) and Apertium with handtagged NEs (nes)
- Test set: nc-2007 (WMT'08)
- Metrics: UNK, BLEU, NIST, TER, GTM

- System: Apertium es–en 0.7.1
- Baselines: Apertium without NEs (no_nes) and Apertium with handtagged NEs (nes)
- Test set: nc-2007 (WMT'08)
- Metrics: UNK, BLEU, NIST, TER, GTM
- Parameters: Min occurrences {25, 50, 100, 200}, min % occurrences capitalised .7, .75, .8, .85

- what is the importance of handtagged NEs in Apertium's dictionaries?

- what is the importance of handtagged NEs in Apertium's dictionaries?
- Apertium with handtagged NEs vs Apertium without NEs

- what is the importance of handtagged NEs in Apertium's dictionaries?
- Apertium with handtagged NEs vs Apertium without NEs

- what is the importance of handtagged NEs in Apertium's dictionaries?
- Apertium with handtagged NEs vs Apertium without NEs

System	UNK	BLEU	NIST	TER	GTM
en→es no_nes	3440	0.1976	6.5389	0.6222	0.4917
en→es nes	2285	0.2119	6.7641	0.6084	0.5054

- what is the importance of handtagged NEs in Apertium's dictionaries?
- Apertium with handtagged NEs vs Apertium without NEs

System	UNK	BLEU	NIST	TER	GTM
en→es no_nes	3440	0.1976	6.5389	0.6222	0.4917
en→es nes	2285	0.2119	6.7641	0.6084	0.5054
es→en no_nes	3027	0.2016	6.1521	0.7091	0.5073
es→en nes	1936	0.2127	6.3277	0.6969	0.5182

Experiments to answer two questions:

1. Can NEs from MINELex obtain comparable performance to handtagged NEs?

Experiments to answer two questions:

1. Can NEs from MINELex obtain comparable performance to handtagged NEs?
2. Can NEs from MINELex add significant value to handtagged NEs?

Adding NEs to Apertium without NEs (en→es)

System	UNK	BLEU	NIST	TER	GTM
100,.75	2334	.2060	6.6879	.6173	.5007
100,.8	2372	**.2061**	6.6882	.6168	**.5010**
200,.75	2441	.2058	**6.6903**	.6164	.5006
200,.8	2481	.2059	6.6899	**.6158**	.5009
no_nes	3440	.1976	6.5389	.6222	.4917
nes	2285	.2119	6.7641	.6084	.5054

Adding NEs to Apertium without NEs (en→es)

System	UNK	BLEU	NIST	TER	GTM
100,.75	2334	.2060	6.6879	.6173	.5007
100,.8	2372	**.2061**	6.6882	.6168	**.5010**
200,.75	2441	.2058	**6.6903**	.6164	.5006
200,.8	2481	.2059	6.6899	**.6158**	.5009
no_nes	3440	.1976	6.5389	.6222	.4917
nes	2285	.2119	6.7641	.6084	.5054

- Automatic NEs >> no_nes for all metrics

Adding NEs to Apertium without NEs (en→es)

System	UNK	BLEU	NIST	TER	GTM
100,.75	2334	.2060	6.6879	.6173	.5007
100,.8	2372	**.2061**	6.6882	.6168	**.5010**
200,.75	2441	.2058	**6.6903**	.6164	.5006
200,.8	2481	.2059	6.6899	**.6158**	.5009
no_nes	3440	.1976	6.5389	.6222	.4917
nes	2285	.2119	6.7641	.6084	.5054

- Automatic NEs >> no_nes for all metrics
- Handtagged NEs >> Automatic NEs for all metrics

Adding NEs to Apertium without NEs (es→en)

System	UNK	BLEU	NIST	TER	GTM
100,.75	2078	**.2100**	**6.2882**	**.7017**	.5152
100,.8	2100	.2099	6.2831	.7019	.5149
200,.75	2303	.2097	6.2826	.7021	.5146
200,.8	2325	.2096	6.2790	.7023	.5144
nones	3027	.2016	6.1521	.7091	.5073
nes	1936	.2127	6.3277	.6969	.5182

Adding NEs to Apertium without NEs (es→en)

System	UNK	BLEU	NIST	TER	GTM
100,.75	2078	**.2100**	**6.2882**	**.7017**	.5152
100,.8	2100	.2099	6.2831	.7019	.5149
200,.75	2303	.2097	6.2826	.7021	.5146
200,.8	2325	.2096	6.2790	.7023	.5144
nones	3027	.2016	6.1521	.7091	.5073
nes	1936	.2127	6.3277	.6969	.5182

- Automatic NEs >> no_nes for all metrics

Adding NEs to Apertium without NEs (es→en)

System	UNK	BLEU	NIST	TER	GTM
100,.75	2078	**.2100**	**6.2882**	**.7017**	.5152
100,.8	2100	.2099	6.2831	.7019	.5149
200,.75	2303	.2097	6.2826	.7021	.5146
200,.8	2325	.2096	6.2790	.7023	.5144
nones	3027	.2016	6.1521	.7091	.5073
nes	1936	.2127	6.3277	.6969	.5182

- Automatic NEs >> no_nes for all metrics
- Handtagged NEs comparable to Automatic (BLEU, NIST)

Adding NEs to Apertium with NEs (en→es)

System	UNK	BLEU	NIST	TER	GTM
25,.8	**2027**	.2105	6.7122	.6144	.5028
100,.75	2089	.2113	6.7472	.6097	.5051
100,.8	2089	.2113	6.7482	.6096	.5052
200,.75	2141	**.2117**	6.7568	.6088	.5054
200,.8	2141	**.2117**	**6.7577**	**.6087**	**.5055**
nes	2285	.212	6.764	.608	.505

Adding NEs to Apertium with NEs (en→es)

System	UNK	BLEU	NIST	TER	GTM
25,.8	**2027**	.2105	6.7122	.6144	.5028
100,.75	2089	.2113	6.7472	.6097	.5051
100,.8	2089	.2113	6.7482	.6096	.5052
200,.75	2141	**.2117**	6.7568	.6088	.5054
200,.8	2141	**.2117**	**6.7577**	**.6087**	**.5055**
nes	2285	.212	6.764	.608	.505

- Automatic+Handtagged NEs comparable to Handtagged

Adding NEs to Apertium with NEs (en→es)

System	UNK	BLEU	NIST	TER	GTM
25,.8	**2027**	.2105	6.7122	.6144	.5028
100,.75	2089	.2113	6.7472	.6097	.5051
100,.8	2089	.2113	6.7482	.6096	.5052
200,.75	2141	**.2117**	6.7568	.6088	.5054
200,.8	2141	**.2117**	**6.7577**	**.6087**	**.5055**
nes	2285	.212	6.764	.608	.505

- Automatic+Handtagged NEs comparable to Handtagged
- UNK can be reduced up to 11.3% (25,.8)

Adding NEs to Apertium with NEs (es→en)

System	UNK	BLEU	NIST	TER	GTM
25,.8	**1725**	.2133	6.3291	.6979	.5184
100,.75	1789	**.2135**	6.3368	**.6968**	**.5188**
100,.8	1789	**.2135**	6.3362	**.6968**	.5187
200,.75	1830	**.2135**	**6.3362**	**.6968**	.5187
200,.8	1830	**.2135**	6.3356	.6969	.5186
nes	1936	.2127	6.3277	.6969	.5182

Adding NEs to Apertium with NEs (es→en)

System	UNK	BLEU	NIST	TER	GTM
25,.8	**1725**	.2133	6.3291	.6979	.5184
100,.75	1789	**.2135**	6.3368	**.6968**	**.5188**
100,.8	1789	**.2135**	6.3362	**.6968**	.5187
200,.75	1830	**.2135**	**6.3362**	**.6968**	.5187
200,.8	1830	**.2135**	6.3356	.6969	.5186
nes	1936	.2127	6.3277	.6969	.5182

- Automatic+Handtagged NEs >> Handtagged

Adding NEs to Apertium with NEs (es→en)

System	UNK	BLEU	NIST	TER	GTM
25,.8	**1725**	.2133	6.3291	.6979	.5184
100,.75	1789	**.2135**	6.3368	**.6968**	**.5188**
100,.8	1789	**.2135**	6.3362	**.6968**	.5187
200,.75	1830	**.2135**	**6.3362**	**.6968**	.5187
200,.8	1830	**.2135**	6.3356	.6969	.5186
nes	1936	.2127	6.3277	.6969	.5182

- Automatic+Handtagged NEs >> Handtagged
- UNK can be reduced up to 10.9% (25,.8)

- Importance of NEs in RBMT (en–es) has been studied

- Importance of NEs in RBMT (en–es) has been studied
  - Improvement across a set of MT evaluation metrics

- Importance of NEs in RBMT (en–es) has been studied
  - Improvement across a set of MT evaluation metrics
  - Reduction by 33% of unknown terms

- Importance of NEs in RBMT (en–es) has been studied
  - Improvement across a set of MT evaluation metrics
  - Reduction by 33% of unknown terms
- Method for enriching RBMT with automatically acquired NEs

- Importance of NEs in RBMT (en–es) has been studied
  - Improvement across a set of MT evaluation metrics
  - Reduction by 33% of unknown terms
- Method for enriching RBMT with automatically acquired NEs
  - System with automatic NEs outperforms system without NEs

- Importance of NEs in RBMT (en–es) has been studied
  - Improvement across a set of MT evaluation metrics
  - Reduction by 33% of unknown terms
- Method for enriching RBMT with automatically acquired NEs
  - System with automatic NEs outperforms system without NEs
  - Mixed results when comparing/adding automatic NEs to handtagged NEs

- Importance of NEs in RBMT (en–es) has been studied
  - Improvement across a set of MT evaluation metrics
  - Reduction by 33% of unknown terms
- Method for enriching RBMT with automatically acquired NEs
  - System with automatic NEs outperforms system without NEs
  - Mixed results when comparing/adding automatic NEs to handtagged NEs
- Software developed

- Importance of NEs in RBMT (en–es) has been studied
  - Improvement across a set of MT evaluation metrics
  - Reduction by 33% of unknown terms
- Method for enriching RBMT with automatically acquired NEs
  - System with automatic NEs outperforms system without NEs
  - Mixed results when comparing/adding automatic NEs to handtagged NEs
- Software developed
  - minelex2plain $\rightarrow$ exports a subset of NEs for a language pair to a plain text tabbed format

- Importance of NEs in RBMT (en–es) has been studied
  - Improvement across a set of MT evaluation metrics
  - Reduction by 33% of unknown terms
- Method for enriching RBMT with automatically acquired NEs
  - System with automatic NEs outperforms system without NEs
  - Mixed results when comparing/adding automatic NEs to handtagged NEs
- Software developed
  - minelex2plain $\rightarrow$ exports a subset of NEs for a language pair to a plain text tabbed format
  - minelex2apertium $\rightarrow$ inserts in Apertium dictionaries output from minelex2plain

- Importance of NEs in RBMT (en–es) has been studied
  - Improvement across a set of MT evaluation metrics
  - Reduction by 33% of unknown terms
- Method for enriching RBMT with automatically acquired NEs
  - System with automatic NEs outperforms system without NEs
  - Mixed results when comparing/adding automatic NEs to handtagged NEs
- Software developed
  - minelex2plain $\rightarrow$ exports a subset of NEs for a language pair to a plain text tabbed format
  - minelex2apertium $\rightarrow$ inserts in Apertium dictionaries output from minelex2plain

- Importance of NEs in RBMT (en–es) has been studied
  - Improvement across a set of MT evaluation metrics
  - Reduction by 33% of unknown terms
- Method for enriching RBMT with automatically acquired NEs
  - System with automatic NEs outperforms system without NEs
  - Mixed results when comparing/adding automatic NEs to handtagged NEs
- Software developed
  - minelex2plain $\rightarrow$ exports a subset of NEs for a language pair to a plain text tabbed format
  - minelex2apertium $\rightarrow$ inserts in Apertium dictionaries output from minelex2plain

`http://www.computing.dcu.ie/~atoral/#Resources`

Thanks! Questions?