Automatic acquisition of Named Entities for Rule-Based Machine Translation*

Antonio Toral
School of Computing
Dublin City University
Ireland
atoral@computing.dcu.ie

Andy Way
School of Computing
Dublin City University
Ireland
away@computing.dcu.ie

Abstract

This paper proposes to enrich RBMT dictionaries with Named Entities (NEs) automatically acquired from Wikipedia. The method is applied to the Apertium English–Spanish system and its performance compared to that of Apertium with and without handtagged NEs. The system with automatic NEs outperforms the one without NEs, while results vary when compared to a system with hand-tagged NEs (results are comparable for Spanish→English but slightly worst for English→Spanish). Apart from that, adding automatic NEs contributes to decreasing the amount of unknown terms by more than 10%.

1 Introduction

NEs usually refer to several types of proper nouns (e.g. people, locations, organisations) and in some cases also to numeric expressions (e.g. data, time, currency). In this work the term NE is used as a synonym of proper noun.

Let us take a look at the distribution of NEs in running text by using the English version of Europarl (Koehn, 2005). This corpus is Part-of-Speech (PoS) tagged and NEs are identified with the FreeLing toolkit (Atserias et al., 2006). The mean number of times each instance is seen in the corpus is very low for NEs (1) compared to other PoS, such as common nouns (3) and verbs (7). Likewise, the average of number of occurrences (24) is also much lower for NEs than for common nouns (295) and verbs (888). Conversely, the number of different instances is much higher (87,682) than for common nouns (26,918) and verbs (7,635).

These distributional properties of NEs (a huge amount of different instances and a very low number of occurrences per instance) together with their dynamic nature (new NEs appear at a much higher rate than for other PoS) (Mann, 2002) make it impractical to build NE resources manually. On the other hand, their morphology is simpler compared to other PoS,[1] and therefore automatic acquisition procedures are more feasible.

The aim of this paper is to add automatically acquired NEs to the dictionaries of a Rule-Based Machine Translation (RBMT) system. Specifically, we consider the Apertium (Tyers et al., 2010) English–Spanish engine. Around one third of the entries (8,000) in its bilingual dictionary are proper nouns. However, they cover less than 10% of the NEs that appear in the English version of Europarl.

The rest of the paper is structured as follows. The following section introduces MINELex, a NE lexicon derived from Wikipedia. After that we introduce our methodology, which basically adds NEs extracted from MINELex to Apertium's dictionaries. This is followed by a description of the software developed. Subsequently, we provide

* This research has been partially funded by the EU project PANACEA (7FP-ITC-248064).

[1] This applies to the languages covered in the article. In some languages NEs are inflected and thus this claim does not apply.

F. Sánchez-Martínez, J.A. Pérez-Ortiz (eds.)
Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, p. 37–43
Barcelona, Spain, January 2011. http://hdl.handle.net/10609/5644

the evaluation, and compare the performance of the new system to vanilla Apertium. Finally we outline some conclusions and propose lines of future work.

2 MINELex

The Multilingual and Interoperable Named Entity Lexicon (MINELex) (Toral et al., 2008; Attia et al., 2010) is a language resource made up of NEs automatically acquired from Wikipedia for 11 languages[2] and connected to semantic units of four computational lexicons (English WordNet (Fellbaum, 1998), Spanish WordNet (Verdejo, 1999), Arabic WordNet (Rodríguez et al., 2008) and the Italian PAROLE-SIMPLE (Ruimy et al., 2002)) and to nodes of two ontologies (SUMO (Niles and Pease, 2001) and SIMPLE (Lenci et al., 2000)). In addition, equivalent NEs in different languages are connected by means of interlingual links. Each NE is associated with confidence scores (the number of occurrences of the NE in a corpus and the percentage of times it occurs capitalised), thus allowing the selection of different subsets of the resource according to the requirements and purpose of the application.

Table 1 summarises the number of NEs, variants of these NEs (different written forms) and relations of these NEs for English and Spanish.

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

Table 1: NEs in MINELex

3 Methodology

Given a language pair, our method extracts pairs of equivalent NEs from MINELex that satisfy certain restrictions (the NE has a minimum number of occurrences and a minimum percentage of occurrences are capitalised). Subsequently, these NEs are inserted into Apertium's dictionaries. This entails inserting the source NE in Apertium's source-language dictionary, the target NE in Apertium's target-language dictionary and transfer information in the bilingual dictionary. No morphology or semantic information is considered. For each NE to be inserted in a monolingual dictionary, these attributes are set: category proper noun (np), subcategory generic (al) and number singular-plural (sp). For Spanish, an additional attribute gender with the value masculine-feminine (mf) is added.

The following is an example of a NE from MINELex and the corresponding entries that are created in the Apertium dictionaries. The element "e" in the XML code contains an entry while the element "pardef" defines an inflection paradigm, which can be shared among several entries.[3]

MINELex data:

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

Apertium English dictionary:

```
<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>
```

Apertium Spanish dictionary:

```
<pardef n="Aquisgrán__np">
 <e><p><l/><r>
  <s n="np"/>
  <s n="al"/>
  <s n="mf"/>
  <s n="sp"/>
 </r></p></e>
</pardef>
[...]
```

[2] Arabic, Catalan, Dutch, English, French, Italian, Norwegian, Portuguese, Romanian, Spanish and Swedish

[3] Detailed information about Apertium's dictionary format can be found in the following URL `http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf`

```
<e lm="Ekaterimburgo">
 <i>Ekaterimburgo</i>
 <par n="Aquisgrán__np"/>
</e>
```

Apertium English–Spanish dictionary:

```
<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>
```

4 Software

Two command-line applications have been developed in order to carry out the methodology presented in the previous section: minelex2plain (m2p) and minelex2apertium (m2a).

m2p is a C++ program that exploits the MINELex API in order to extract a subset of NEs and related data according to a set of parameters. These are the source-language, the target-language, a threshold for the minimum number of occurrences, a threshold for the minimum percentage of capitalised occurrences, whether to ignore those NEs whose lemmas are equal in both languages and whether to ignore variants (i.e. only output full forms). The output of this program is plain text where each line contains a NE equivalence between the two languages and is made up of a set of fields separated by tabs: NE in the source-language, NE in the target language, direction[4] (LRL if both NEs are full forms, LR if the source NE is a variant and RL if the target NE is a variant), number of occurrences and percentage of capitalised occurrences. An example follows:

```
Yekaterinburg   Ekaterimburgo  LRL 190 .95
Ekaterinenburg Ekaterimburgo  LR  190 .95
Yekaterinburg   Yekaterimburgo RL  190 .95
```

m2a is a perl script that reads as input the output from m2p and inserts the relevant data into Apertium dictionaries. It takes as parameters three Apertium dictionary files (source-language, target-language and bilingual dictionary). An example of the output has been already shown in the previous section.

5 Evaluation

This section presents the evaluation. First, in 5.1 we describe the experimental environment. Then, in 5.2, we show the results obtained and draw conclusions from them.

5.1 Data and metrics

The baselines are based on the last stable version of Apertium English–Spanish at the time of writing (0.7.1).[5] Two baselines are considered. The first is the Apertium engine without any modification (en–es nes), while the second is the Apertium engine without NEs (en–es no_nes).

NE-enriched systems are built with different values for the thresholds minimum of occurrences (25, 50, 100 and 200) and minimum percentage of occurrences capitalised (.75, .8, and .85). These values are chosen empirically.

We evaluate the systems on the News Commentary 2007 English–Spanish test set (nc-2007) from WMT08,[6] which contains 2,000 sentence pairs. The following metrics are used in our experiments:, BLEU (Papineni et al., 2002), NIST (Doddington, 2002), TER (Snover et al., 2006), GTM (Turian et al., 2003), METEOR (Lavie and Denkowski, 2009),[7] METEOR-Next (Denkowski and Lavie, 2010)[7] and DCU-LFG (He et al., 2010).[7] Furthermore, we provide for each system execution the amount of unknown tokens (UNK) in the source side of the test set. Statistical significance tests are carried out for BLEU and NIST (with ARK's code)[8] and for GTM (using FastMtEval).[9] P-value is set

[4] This field determines the direction of translation in Apertium. LRL entries will be translated bidirectionally, LR only left to right, and RL only right to left.

[5] http://sourceforge.net/projects/apertium/files/apertium-en-es/0.7/apertium-en-es-0.7.1.tar.gz

[6] http://www.statmt.org/wmt08/devsets.tgz

[7] These are only applied when the target language is English.

[8] http://www.ark.cs.cmu.edu/MT/

[9] http://www.computing.dcu.ie/

to 0.05.

5.2 Experiments

Prior to running the actual experiments, we need to know what is the importance of handtagged NEs in Apertium's dictionaries, so we compare the performance of Apertium with and without NEs.

Results are shown in Table 2. The addition of NEs reduces by roughly one third the number of unknown terms. There is also a notable improvement in performance across all the MT metrics (more than one absolute point for BLEU, TER, GTM and METEOR).

Once the importance of NEs has been demonstrated, we design two experiments in order to provide answers to the following two research questions:

1. Can the NEs from MINELex obtain comparable performance to the handtagged NEs in Apertium's dictionaries?

2. Can the NEs from MINELex add significant value to the handtagged NEs in Apertium's dictionaries?

In the first experiment we add NEs from MINELex to Apertium without NEs and compare the results both to Apertium with NEs and Apertium without NEs. Results are shown in Table 3 (for English→Spanish) and Table 4 (for Spanish→English). The two values in the column System stand for the values for the thresholds (minimum occurrences and minimum percentage of occurrences capitalised).

For both directions, the results obtained are very similar for all the different combinations of the two parameters used in the system. For English→Spanish (Table 3) and for all the metrics, Apertium with handtagged NEs obtains better scores than Apertium with automatic NEs, and this system obtains better scores than Apertium without NEs (absolute improvements of .0085 for BLEU, .15 for NIST, .0064 for TER and .0093 for GTM). In both cases the difference is statistically significant for BLEU, NIST and GTM.

System	UNK	BLEU	NIST	TER	GTM
25,.75	**2150**	.2056	6.6537	.6222	.4985
25,.8	2189	.2057	6.6547	.6217	.4988
25,.85	2276	.2053	6.6396	.6224	.4978
50,.75	2198	.2060	6.6764	.6189	.5000
50,.8	2237	**.2061**	6.6767	.6185	.5003
50,.85	2322	.2056	6.6610	.6192	.4993
100,.75	2334	.2060	6.6879	.6173	.5007
100,.8	2372	**.2061**	6.6882	.6168	**.5010**
100,.85	2457	.2057	6.6723	.6176	.5000
200,.75	2441	.2058	**6.6903**	.6164	.5006
200,.8	2481	.2059	6.6899	**.6158**	.5009
200,.85	2563	.2055	6.6740	.6166	.4999
no_nes	3440	.1976	6.5389	.6222	.4917
nes	2285	.2119	6.7641	.6084	.5054

Table 3: Adding NEs to Apertium without NEs (en→es)

In the case of Spanish→English (Table 4), Apertium with handtagged NEs obtains better scores than Apertium with automatic NEs, while this system obtains better scores than Apertium without NEs (absolute improvements of .0084 for BLEU, .14 for NIST, .0074 for TER, .008 for GTM, .009 for METEOR, .0082 for METEOR-Next and 0.123 for DCU-LFG). In both cases the difference is statistically significant for GTM, whereas only the difference between the last two is statistically significant for BLEU (.21 vs .2016) and NIST (6.2882 vs 6.1521). According to these tests, Apertium Spanish→English with handtagged NEs and with automatic NEs has a comparable performance with respect to BLEU (.21 vs .2127) and NIST (6.2882 vs 6.3277).

In the second experiment we add NEs from MINELex to Apertium with NEs and compare the results to Apertium with NEs. Results are shown in Table 5 for the direction English→Spanish, and Table 6 for the direction Spanish→English.

In this second experiment the results obtained are again very similar for all the different combinations of the two parameters used in the system. For English→Spanish (Table 5), all the MT metrics but GTM exhibit a very slight decrease in performance, although the differences in BLEU (.2117 vs .212), NIST (6.7577 vs 6.764) or GTM (.5055 vs .505) are not statistically significant. On the other hand the amount of unknown terms is reduced by up to 11.3% (system configurations

~nstroppa/index.php?page=softwares

System	UNK	BLEU	NIST	TER	GTM	MET	MET-N	DCU-LFG
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	0.6034	0.5216	0.4970
es→en nes	1936	0.2127	6.3277	0.6969	0.5182	0.6169	0.5315	0.5109

Table 2: Apertium performance with and without NEs

System	UNK	BLEU	NIST	TER	GTM	MET	MET-N	DCU-LFG
25,.75	**1979**	**.2100**	6.2842	.7027	.5151	.6119	.5294	.5084
25,.8	2002	.2098	6.2791	.7029	.5148	.6115	.5292	.5079
25,.85	2087	.2083	6.2584	.7041	.5136	.6102	.5281	.5057
50,.75	2019	**.2100**	6.2879	.7020	**.5153**	**.6124**	**.5298**	**.5093**
50,.8	2042	.2098	6.2828	.7022	.5150	.6119	.5295	.5088
50,.85	2127	.2083	6.2620	.7034	.5137	.6106	.5285	.5066
100,.75	2078	**.2100**	**6.2882**	**.7017**	.5152	.6123	.5297	.5090
100,.8	2100	.2099	6.2831	.7019	.5149	.6118	.5295	.5085
100,.85	2181	.2083	6.2616	.7031	.5136	.6105	.5284	.5063
200,.75	2303	.2097	6.2826	.7021	.5146	.6118	.5295	.5088
200,.8	2325	.2096	6.2790	.7023	.5144	.6114	.5293	.5082
200,.85	2403	.2083	6.2613	.7032	.5134	.6103	.5284	.5064
nones	3027	.2016	6.1521	.7091	.5073	.6034	.5216	.4970
nes	1936	.2127	6.3277	.6969	.5182	.6169	.5315	.5109

Table 4: Adding NEs to Apertium without NEs (es→en)

System	UNK	BLEU	NIST	TER	GTM	MET	MET-N	DCU-LFG
25,.75	**1725**	.2133	6.3297	.6978	.5184	.6172	.5317	.5113
25,.8	**1725**	.2133	6.3291	.6979	.5184	.6172	.5317	.5113
25,.85	1733	.2132	6.3280	.6979	.5183	.6171	.5317	.5112
50,.75	1750	.2134	6.3352	.6970	.5188	**.6178**	**.5322**	**.5122**
50,.8	1750	.2134	6.3346	.6971	.5187	.6177	**.5322**	**.5122**
50,.85	1758	.2133	6.3335	.6971	.5187	.6176	.5321	.5121
100,.75	1789	**.2135**	6.3368	**.6968**	**.5188**	**.6178**	**.5322**	.5118
100,.8	1789	**.2135**	6.3362	**.6968**	.5187	.6177	**.5322**	.5118
100,.85	1793	.2133	6.3344	.6969	.5186	.6176	.5321	.5117
200,.75	1830	**.2135**	**6.3362**	**.6968**	.5187	.6176	.5321	**.5122**
200,.8	1830	**.2135**	6.3356	.6969	.5186	.6175	.5321	**.5122**
200,.85	1831	**.2135**	6.3356	.6969	.5186	.6175	.5321	**.5122**
nes	1936	.2127	6.3277	.6969	.5182	.6169	.5315	.5109

Table 6: Adding NEs to Apertium with NEs (es→en)

System	UNK	BLEU	NIST	TER	GTM
25,.75	**2027**	.2104	6.7112	.6146	.5028
25,.8	**2027**	.2105	6.7122	.6144	.5028
25,.85	2031	.2106	6.7129	.6143	.5029
50,.75	2052	.2108	6.7347	.6115	.5043
50,.8	2052	.2109	6.7357	.6114	.5044
50,.85	2054	.2109	6.7359	.6114	.5044
100,.75	2089	.2113	6.7472	.6097	.5051
100,.8	2089	.2113	6.7482	.6096	.5052
100,.85	2091	.2114	6.7484	.6096	.5052
200,.75	2141	**.2117**	6.7568	.6088	.5054
200,.8	2141	**.2117**	**6.7577**	**.6087**	**.5055**
200,.85	2141	**.2117**	**6.7577**	**.6087**	**.5055**
nes	2285	.212	6.764	.608	.505

Table 5: Adding NEs to Apertium with NEs (en→es)

25,.75 and 25,.8).

Conversely, in the case of English→Spanish (Table 6), we see a slight improvement for all MT metrics, which is statistically significant for BLEU (.2135 vs .2127), NIST (6.3362 vs 6.3277) and GTM (.5122 vs .5109). The number of unknown terms is reduced by up to 10.9% (system configurations 25,.75 and 25,.8).

6 Conclusions

This paper has studied the importance of NEs in the Apertium RBMT system for the English–Spanish language pair and has explored the enrichment of its dictionaries with automatically acquired NEs.

The role of the handtagged NEs in the system has been found to be very relevant as it not only reduces by one third the number of unknown terms, but also exhibits a sustained improvement across a set of MT evaluation metrics. For example, NEs prevent wrong translations in those cases in which the surface form may have other analysis. E.g. the Spanish lemma "Zapatero" might refer to the common noun shoemaker or to the NE president of Spain, therefore identifying the correct meaning in the text is vital in order to produce the correct English translation.

Automatically added NEs improve a system without handtagged NEs. When comparing automatically added NEs to the system with handtagged NEs, different results are found depending on the language direction. For English→Spanish, automatically added NEs perform slightly worse than a system with handtagged NEs. Conversely, for Spanish→English, automatic NEs obtain comparable results to using handtagged NEs.

For English→Spanish, adding NEs to the handtagged ones obtains comparable results, while for the other direction results do significantly increase. In both cases, the addition of NEs reduces the amount of unknown terms by more than 10%.

Yet another contribution of this work is the availability of the software developed under the GNU General Public License. This comprises mainly software that extracts NEs from MINELex and inserts them into Apertium's dictionaries and is available at `http://www.computing.dcu.ie/~atoral/#Resources`.

Regarding future work, we plan to apply this methodology to other language pairs in order to see whether the same trends apply. In addition, we would like to extend the method by acquiring morphologic features, as this would allow to apply our methodology to languages where NEs are inflected.

References

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., and Padró, M. (2006). Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 48–55.

Attia, M., Toral, A., Tounsi, L., Monachini, M., and van Genabith, J. (2010). An automatically built named entity lexicon for arabic. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Denkowski, M. and Lavie, A. (2010). METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. In *Proceedings of the ACL 2010*

Joint Workshop on Statistical Machine Translation and Metrics MATR.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

He, Y., Du, J., Way, A., and van Genabith, J. (2010). The dcu dependency-based metric in wmt-metricsmatr 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 349–353, Uppsala, Sweden. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Lavie, A. and Denkowski, M. J. (2009). The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., and Zampolli, A. (2000). Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

Mann, G. (2002). Fine-grained proper noun ontologies for question answering. In *Proceedings of SemaNet'02: Building and Using Semantic Networks*.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems*, pages 2–9, New York, NY, USA. ACM.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Rodríguez, R., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Mart., M., Black, W., Elkateb, S., Kirk, J., Pease, A., Vossen, P., and Fellbaum, C. (2008). Arabic wordnet: Current state and future extensions. In *The Fourth Global WordNet Conference*, Szeged, Hungary.

Ruimy, N., Monachini, M., Distante, R., Guazzini, E., Molino, S., Ulivieri, M., Calzolari, N., and Zampolli, A. (2002). Clips, a multi-level italian computational lexicon: A glimpse to data. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas de Gran Canaria, Spain.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Toral, A., Muñoz, R., and Monachini, M. (2008). Named entity wordnet. In (ELRA), E. L. R. A., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Turian, J., Shen, L., and Melamed, I. D. (2003). Evaluation of machine translation and its evaluation. In *In Proceedings of MT Summit IX*, pages 386–393.

Tyers, F. M., Sánchez-Martínez, F., Ortiz-Rojas, S., and Forcada, M. L. (2010). Free/open-source resources in the apertium platform for machine translation research and development. *The Prague Bulletin of Mathematical Linguistics*, (93):67—76. ISSN: 0032-6585.

Verdejo, M. F. (1999). The Spanish Wordnet, EuroWordNet Deliverable D032D033 part B3. Technical report.