The UPC System for Arabic-to-English Entity Translation

D. Farwell, J. Gimenez, E. Gonzalez, R. Halkoum, H. Rodriguez, M. Surdeanu

Technical University of Catalonia
{farwell, jgimenez, egonzalez, halkoum, horacio, surdeanu}@lsi.upc.edu

March 30, 2007

Outline

Architecture

Named Entity Recognition and Classification

Coreference Resolution

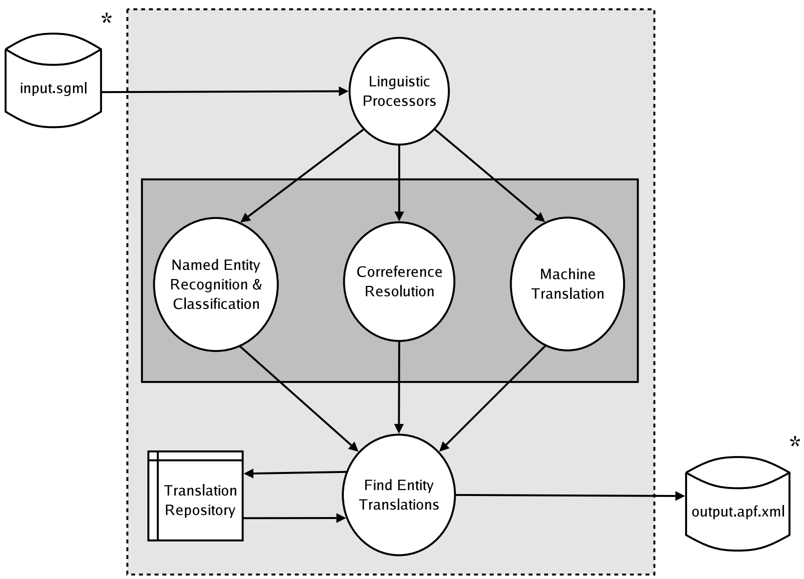
Machine Translation

Resources

Evaluation

Conclusions

System Architecture



Entity Translation (ET) $\equiv$ *disambiguation problem solved through statistical Machine Translation (MT).*

Execution flow:

1. Preprocessing at shallow syntax level.
2. Entity mentions recognized in source Arabic text.
3. Coreference chains extracted in source text.
4. Whole source text translated to English using a statistical phrase-based MT system.
5. Phrases corresponding to entity mentions identified in translation.
6. Mentions merged into entities based on the coreference chains of source text.

Exceptions

- *Untranslated entities*: translation fails (unknown words, unknown context). Solution:
  1. Lookup in the Translation Repository, which contains all entities previously translated.
  2. If no candidate found, inspect the bilingual gazetteer.
  3. If no translation found, output the incomplete translation from the MT system.
- *Phrase boundaries*: because our MT is phrase-based it may happen that an entity mention does not match exactly with a phrase. Solution: output the translation for the text that contains the source entity mention.

Outline

Approach: Sequential BIO Tagger

Input: A training sample $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^m$
Input: Number of epochs T

$\mathbf{M} = \mathbf{0}$ ($\mathbf{M} \in I\!R^{|\mathcal{Y}| \times |\mathcal{X}|}$)
for $t = 1$ **to** T **do**
  for $i = 1$ **to** m **do**
    predict $\hat{y}_i = \arg\max_{r=1}^{|\mathcal{Y}|}\{\langle \mathbf{M}_r, \mathbf{x}_i \rangle\}$
    set $E = \{r \neq y_i : \langle \mathbf{M}_r, \mathbf{x}_i \rangle \geq \langle \mathbf{M}_{y_i}, \mathbf{x}_i \rangle\}$
    if $E \neq \phi$ **then**
      for all r **in** E **do**
        $\mathbf{M}_r = \mathbf{M}_r - \mathbf{x}_i / |E|$
      end for
      $\mathbf{M}_{y_i} = \mathbf{M}_{y_i} + \mathbf{x}_i$
    end if
  end for
end for

Output: $H(\mathbf{x}) = \arg\max_r \{\langle \mathbf{M}_r, \mathbf{x} \rangle\}$

- Learning algorithm: Ultraconservative Multiclass Perceptron Algorithm (UMPA)
  - Maintains a prediction matrix **M** with one row for each class to be modeled.
  - Ultraconservative, it updates only the vectors of the classes that scored higher than the correct class.
- Greedy inference: for every token select the label with the highest score that is consistent with the previous labels.
- Two classifiers trained: one for entity type + subtype (89 classes), another for the entity mention type (`NOM`, `NAM`, `PRO`).

Features

Model M_1 - adds lexical attributes:

- The token lexem.
- The suffixes and prefixes of length 2, 3, and 4.
- The sequence obtained by removing all letters from the token.
- The sequence obtained by removing all alphanumeric characters from the token.
- *isAllDigits* - Boolean flag set to true if the word contains only digits.
- *isAllDigitsOrDots* - Boolean flag set to true if the word contains only digits or dots.

Model M_2 - adds part of speech (POS) attributes.

Model M_3 - adds syntactic chunk labels.

Model M_4 - adds class and gazetteer-based attributes:

- *isNumber* - true if the token is a word-spelled number.
- *isMultiplier* - true if the token is a multiplier typically used to compose numbers.
- *isDay* - true if the token is the name of a day of the week.
- *isMonth* - true if the token is the name of a month.
- *isPersonTrigger* - indicates if the token begins or is inside a person trigger.
- *knownPerson* - indicates if the token is part of a sequence that is an known person name.

All models - static context (preceding/following tokens); dynamic context (previous labels).

Evaluation

- Training: ACE 2005 + 2007 (780 docs); development (19 docs).

Model	P	R	F_1	Best epoch
M1	76.54%	75.27%	75.90	15
M2	76.43%	77.32%	76.87	18
M3	77.51%	**77.81%**	**77.66**	19
M4	**79.91%**	70.38%	74.84	29

NERC results on the development set for the entity type/subtype problem.

Model	P	R	F_1	Best epoch
M1	78.25%	78.79%	78.52	31
M2	78.54%	**79.77%**	**79.15**	35
M3	78.30%	79.37%	78.83	35
M4	**80.20%**	69.70%	74.58	35

NERC results on the development set for the entity mention type problem.

- M3 best for entity type $+$ subtype; M2 best for mention type.
- Quantitative analysis: training time 175 seconds/epoch. Labels 1,600 words/second.

Outline

Approach 1: Round Robin Resolution

Input: A text T

for all Pronouns p in T **do**
  Find candidate set C
  Filter candidate set
   $C' = \{c \in C \mid F_1(p,c) > 0\}$
  if C' is empty **then**
    Pronoun p is considered unsolved
  else
    Initialize scores $\forall c \in C'$ $score[c] = 0$
    for all Pairs $c_1, c_2 \in C'$ where
     $dist(c_1, p) < dist(c_2, p)$ **do**
      if $F_2(p, c_1, c_2) > 0$ **then**
        Increment $score[c_1]$
      else
        Increment $score[c_2]$
      end if
    end for
    Set $c_a = \arg\max_c score[c]$ as the
     antecedent of p
  end if
end for

Output: The text T with pronouns resolved

Is candidate **X** *a better antecedent of pronoun* **P** *than candidate* **Y***?*

Execution flow:

1. Construct the set of all candidates that pass the filter F_1.
2. Compare each candidate with the others (F_2). Increment score of best candidate.
3. Select candidate with the highest score.

Approach 2: Lineal Resolution

```
Input: A text T
  for all Pronouns p in T do
    Find candidate set C
    Filter candidate set
      C' = {c ∈ C | F1(p, c) > 0}
    if C' is empty then
      Pronoun p is considered unsolved
    else
      Set as best candidate c_b the candidate in C'
      closest to p
      for all Candidates c ∈ C'
        from closest to furthest to p do
         if F2(p, c_t, c) < 0 then
           Set c as new best candidate c_b
         end if
      end for
      Set the best candidate c_b as the
        antecedent of p
    end if
  end for
Output: The text T with pronouns resolved
```

Is candidate **X** *a better antecedent of pronoun* **P** *than candidate* **Y**?

Execution flow:

1. Construct the set of all candidates that pass the filter F_1.
2. Set as best candidate the closest to the pronoun.
3. Inspect all candidates from closest to furthest to the pronoun. Greedily update the best candidate.

Details

Features:

- Language independent features: form, POS tag, and chunk tag for pronoun, candidate, and a given context window for both pronoun and candidate.
- Language dependent features:
  - Flag that indicates if ASVM-Tools had to change the word form to restore the feminine marker (simple indicator of genre).
  - The word starts with the determinant *Al*.

Classifier:

Support Vector Machines with a polynomial kernel of degree 2.

Evaluation

- Corpus: the Newswire section of ACE 2005 + 2007. Training: 453 documents; development: 40 docs.
- Candidate search span: current sentence $+$ 2 previous sentences. Context window size: $\pm$5 words.

Model		Overall Assignation	Evaluable Assignation	Evaluable Precision	Evaluable Recall
Round Robin	Filter	46%	52%	**65%**	34%
	No	100%	100%	11%	11%
Lineal	Filter	46%	52%	63%	33%
	No	100%	100%	50%	**50%**

Coreference resolution performance.

Training	F_1	6h 8min
	F_2	167h 39min

Round Robin	Filter	7min
	No	4h 55min
Lineal	Filter	7min
	No	26min

Quantitative analysis.

- Precision more important: selected the Round Robin algorithm with filtering.

Outline

Approach

- Phrase-based statistical MT built using freely-available components.
- Trigram language models built using the *SRI Language Modeling Toolkit*.
- Translation models built using word-aligned corpora.
  - Word alignments generated with *GIZA++ SMT Toolkit*.
  - The phrase-extract algorithm of Och (2002) applied on the Viterbi output of Giza++. Considered phrases up to length 5. Phrase pairs scored using unsmoothed Maximum Likelihood Estimation (MLE).
  - The *Pharaoh* beam search decoder used for the `arg max` search. Probability models combined in a log-linear fashion:

$$
\begin{aligned}
logP(e|f) \propto\; & \lambda_{lm1} logP(e)_1 + \ldots + \lambda_{lmN} logP(e)_M \\
& + \lambda_{fe1} logP(f|e)_1 + \ldots + \lambda_{feN} logP(f|e)_N \\
& + \lambda_{ef1} logP(e|f)_1 + \ldots + \lambda_{efN} logP(e|f)_N
\end{aligned}
$$

Experimental Settings

Translation models:

- AE Arabic English Parallel News.
- AR Arabic News Translation Text.
- UN United Nations (2000-2002). For practical reasons we limit to the portion covering years 2000-2002 (1,339,339 sentence pairs, 50.3 million Arabic words, 45.5 million English words).

English language models:

- AE Arabic English Parallel News.
- AR Arabic News Translation Text.
- AM ACE 2005 Multilingual Training Corpus.
- AU ACE 2005 Multilingual Unsupervised Training Data.
- UN United Nations (1993-2002).

System parameters tuned to maximize the overlap of named entities between translation and reference.

Evaluation

- Two development corpora used: DEV_{AE} consists of 961 sentence pairs extracted from the 'AE' corpus (in domain); DEV_{ET} is based on a subset of 987 sentence pairs from the 'REFLEX' training and development set.

metric	**DEV**$_{AE}$	**DEV**$_{ET}$
BLEU-4	0.19	0.06
GTM-1	0.17	0.12
MTR-wnsyn	0.56	0.23
NIST-5	5.55	2.65
RG-W-1.2	0.23	0.15
NE-overlap-**	0.30	0.12
NE-match-*	0.37	0.10

MT performance

- Quantitative analysis: training on the AE corpus – 1 day; training on the UN corpus – almost 3 weeks. Translation time: 72 seconds/document (includes preprocessing).

Outline

Resources

Gazetteers:

All gazetteers used in our system belong to *BADR (Barcelona Arabic Database for Named Entity Recognition)*. Contains:

- BARTIme: temporal expressions.
- BARMOney: monetary expressions.
- BARNAme: names of people.
- BARCO: organizations, associations, names of companies.
- BARLO: locations, cities, districts.

Tools:

Linguistic Processing of Arabic performed using the ASVM-Tools: sentences are transformed into Buckwalter's encoding, tokenized, lemmatized, part-of-speech (PoS) tagged, and base phrase chunked.

Language models are built using the *SRI Language Modeling Toolkit*.

Word alignments are obtained using the *GIZA++ SMT Toolkit*.

Outline

Overall Results

- Out of 4189 entities:
  - Identified correctly: 853 (20.36%);
  - Partially identified: 1089 (26.00%);
  - Failed to identify: 2247 (53.64%);
  - False positives: 3421.
- Proper nouns (43.73%):
  - Identified correctly: 499 (27.24%);
  - Partially identified: 389 (21.23%);
  - Failed to identify: 944 (51.53%);
  - False positives: 1630.
- Common nouns (49.96%):
  - Identified correctly: 355 (16.96%);
  - Partially identified: 623 (29.77%);
  - Failed to identify: 1115 (53.27%);
  - False positives: 1760.
- Pronouns (6.30%):
  - Identified correctly: 8 (3.03%);
  - Partially identified: 68 (25.76%);
  - Failed to identify: 188 (71.21%);
  - False positives: 311.

Diagnostic Results

- Out of 4189 entities:
  - Identified correctly: 1066 (25.45%);
  - Partially identified: 1068 (25.50%);
  - Failed to identify: 2055 (49.05%);
  - False positives: 4635 (we used predicted coreference chains!)
- Proper nouns (43.73%):
  - Identified correctly: 627 (34.22%);
  - Partially identified: 318 (17.36%);
  - Failed to identify: 887 (48.42%);
  - False positives: 1917.
- Common nouns (49.96%):
  - Identified correctly: 433 (20.69%);
  - Partially identified: 667 (31.87%);
  - Failed to identify: 993 (47.44%);
  - False positives: 2365.
- Pronouns (6.30%):
  - Identified correctly: 6 (2.27%);
  - Partially identified: 83 (31.44%);
  - Failed to identify: 175 (66.29%);
  - False positives: 353.

Other Common Errors

- Manually analyzed 498 errors that are not coreference errors nor complete MT mistakes.
- Error distribution:
  - 274 (55.02%) were misidentified named entities, e.g., الَانفصَالية هَى المسءولة (*"separatism is responsible of"*) tagged as NE. Out of these 115 (23.09%) caused by poor stemming (the determinant *Al* not separated → yields spurious NEs).
  - 132 (26.51%) were mistranslated, e.g., ايرلَاندَا الشمَالية translated as *"a a"*, should be *"North Ireland"*.
  - 38 (7.63%) were partially translated, e.g., الَاسلحة translated as *"of weapons"*, *"of"* has been wrongly added.
  - The others are NERC errors, e.g., partially identified entities or misclassified entities.

Conclusions

- Proposed a complete ET model where all components modeled with machine learning. The system core based on statistical MT.
- Overall results not so good (solid -60 value score, but decent unweighted F score). But this is a baseline system.
- Large room for improvement:
  - NER: process destination language (LDC's perfect matching Arb-Eng: 62.3%).
  - NER: generate extent?
  - MT: train on data from ACE domains.
  - MT: change to discriminative specialized models that focus on entity translation.
  - CR: (a) handle non-pronominal coreference; (b) handle cataphora; and (c) better features (tuned for ACE).
  - Output format: generate the NAME attributes.
  - Better component integration. Joint NER $+$ MT model?
  - Talk to each other (the NAME attributes, the *Al* bug)...
- Our approach is (largely) language-independent $\rightarrow$ address other languages as future work.

Thank you! Questions?