Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited

Miquel Esplà-Gomis Felipe Sánchez-Martínez
Mikel L. Forcada
{mespla,fsanchez,mlf}@dlsi.ua.es

Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain

15th Annual Conference of the EAMT
Leuven, May 30, 2011

Outline

1. Introduction
2. Related Work
3. Methodology
4. Experiments and Results
5. Conclusion
6. Current and future Work

Outline

1. Introduction
2. Related Work
3. Methodology
4. Experiments and Results
5. Conclusion
6. Current and future Work

Translation Memories

English	Catalan
s_1: European Association for Machine Translation	t_1: Associació Europea per a la Traducció Automàtica
s_2: The EAMT is a member of the IAMT	t_2: L'EAMT és membre de l'IAMT
s_3: current year's conference is held in Leuven	t_3: el congrés d'enguany se celebra a Lovaina
...	...

Translation Memories

English	Catalan
s_1: European Association for Machine Translation	t_1: Associació Europea per a la Traducció Automàtica
s_2: The EAMT is a member of the IAMT	t_2: L'EAMT és membre de l'IAMT
s_3: current year's conference is held in Leuven	t_3: el congrés d'enguany se celebra a Lovaina
...	...

New sentence s': The AMTA is a member of the IAMT

Translation Memories

English	Catalan
s_1: European Association for Machine Translation	t_1: Associació Europea per a la Traducció Automàtica
s_2: The EAMT is a member of the IAMT	t_2: L'EAMT és membre de l'IAMT
s_3: current year's conference is held in Leuven	t_3: el congrés d'enguany se celebra a Lovaina
...	...

New sentence s': The AMTA is a member of the IAMT

Best match s_2: The EAMT is a member of the IAMT

Translation Memories

English	Catalan
s_1: European Association for Machine Translation	t_1: Associació Europea per a la Traducció Automàtica
s_2: The EAMT is a member of the IAMT	t_2: L'EAMT és membre de l'IAMT
s_3: current year's conference is held in Leuven	t_3: el congrés d'enguany se celebra a Lovaina
...	...

New sentence	s':	The AMTA is a member of the IAMT
Best match	s_2:	The EAMT is a member of the IAMT
Proposal	t_2:	L'EAMT és membre de l'IAMT

Fuzzy Matching Scores

Fuzzy matching scores measure the similarity between segments s' (segment to be translated) and s_i (matching segment in the Translation memory)

$$\mathrm{score}(s', s_i) = 1 - \frac{\mathrm{EditDistance}(s', s_i)}{\max(|s'|, |s_i|)}$$

Fuzzy Matching Scores

Fuzzy matching scores measure the similarity between segments s' (segment to be translated) and s_i (matching segment in the Translation memory)

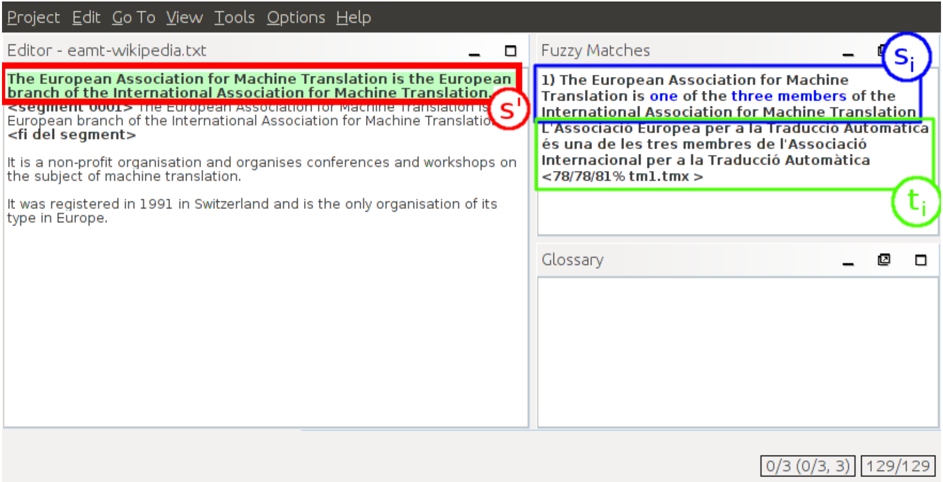
$$\mathrm{score}(s', s_i) = 1 - \frac{\mathrm{EditDistance}(s', s_i)}{\max(|s'|, |s_i|)}$$

Example

s': The Association for Machine Translation in the Americas is the American branch of the IAMT
s_i: The European Association for Machine Translation is a member of the IAMT

$$\mathrm{score}(s', s_i) = 1 - \frac{7}{15} \simeq 0,53$$

Translation-Memory Based CAT Tools


Project Edit Go To View Tools Options Help
Editor - eamt-wikipedia.txt
Fuzzy Matches
Glossary
s'
s_i
t_i

The European Association for Machine Translation is the European branch of the International Association for Machine Translation.

<segment 0001> The European Association for Machine Translation is the European branch of the International Association for Machine Translatio
<fi del segment>

It is a non-profit organisation and organises conferences and workshops on the subject of machine translation.

It was registered in 1991 in Switzerland and is the only organisation of its type in Europe.

1) The European Association for Machine Translation is one of the three members of the International Association for Machine Translation

L'Associació Europea per a la Traducció Automàtica és una de les tres membres de l'Associació Internacional per a la Traducció Automàtica

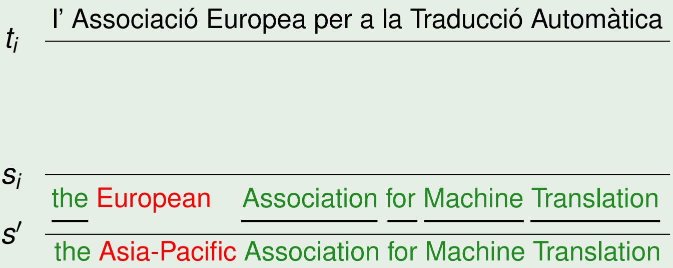
<78/78/81% tm1.tmx >

0/3 (0/3, 3) 129/129

Fuzzy Match Scores + Alignment

Edit distance provides information about the matching words between s' and s_i:

Example



t_i l' Associació Europea per a la Traducció Automàtica

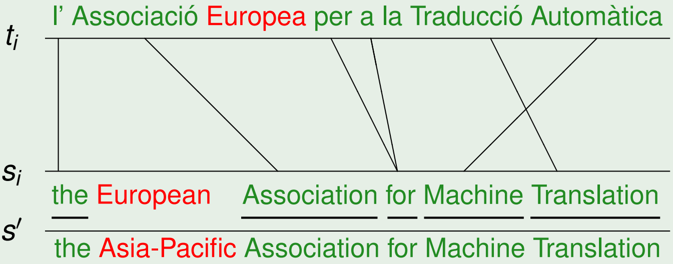
s_i the European Association for Machine Translation

s' the Asia-Pacific Association for Machine Translation

Fuzzy Match Scores + Alignment

Word alignment may be used to “project” source-side matching information onto t_i to suggest which words to change and which to keep unedited:

Example


t_i
l’ Associació Europea per a la Traducció Automàtica
s_i
the European Association for Machine Translation
s'
the Asia-Pacific Association for Machine Translation

Outline

1. Introduction
2. **Related Work**
3. Methodology
4. Experiments and Results
5. Conclusion
6. Current and future Work

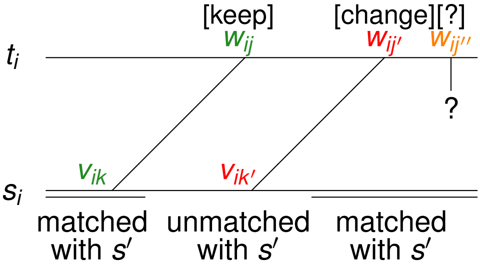
Related Work

- **Simard (2003)**: Statistical MT techniques allows exploiting TMs at sub-segment (sub-sentential) level: *translation spotting*
- **Bourdaillet et al. (2009)**: Similar approach for a bilingual concordancer, *TransSearch*
- **Kranias and Samiotou (2004)**: Sub-segment level alignments using a bilingual dictionary to (i) detect words to be changed and (ii) propose translations for them

Outline

1. Introduction
2. Related Work
3. Methodology
4. Experiments and Results
5. Conclusion
6. Current and future Work

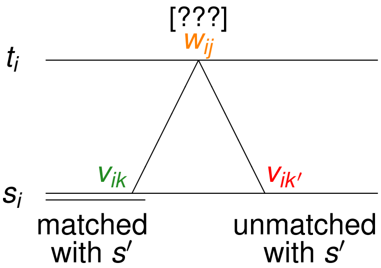
Rationale



- w_{ij} and v_{ik} **aligned** and v_{ik} **matched** $\Longrightarrow$ **keep** w_{ij}
- w_{ij} and v_{ik} **aligned** and v_{ik} **not matched** $\Longrightarrow$ **change** w_{ij}
- w_{ij} **not aligned** $\Longrightarrow$ **???**

Rationale

What to do if there is more than one alignment with contradictory evidence?



Rationale

We define the likelihood of keeping the word w_{ij} unedited as:

$$f_K(w_{ij}, s', s_i, t_i) = \frac{\sum_{v_{ik} \in \mathrm{aligned}(w_{ij})} \mathrm{matched}(v_{ik})}{|\mathrm{aligned}(w_{ij})|}$$

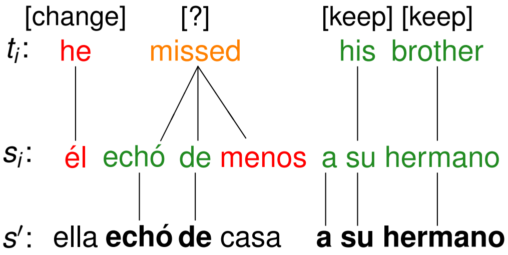
- $\mathrm{aligned}(w_{ij})$: set of source-side words aligned with w_{ij} in s_i
- $\mathrm{matched}(v_{ik})$: 1 if v_{ik} is matched in s' and 0 otherwise

Interpretation of $f_K(w_{ij}, s', s_i, t_i)$

Two ways to interpret $f_K(w_{ij}, s', s_i, t_i)$:

- Unanimity:
  - if $f_K(w_{ij}, s', s_i, t_i) = 1$: w_{ij} →keep unedited
  - if $f_K(w_{ij}, s', s_i, t_i) = 0$: w_{ij} → change
  - otherwise → not marked
- Majority:
  - if $f_K(w_{ij}, s', s_i, t_i) > \frac{1}{2}$: w_{ij} → keep unedited
  - if $f_K(w_{ij}, s', s_i, t_i) < \frac{1}{2}$: w_{ij} → change
  - otherwise → not marked

Example of Unanimity Criterion



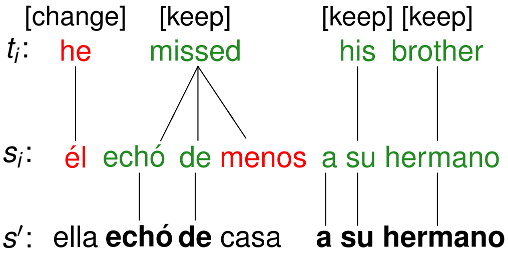
[change] [?] [keep] [keep]

t_i: he missed his brother

s_i: él echó de menos a su hermano

s': ella **echó** **de** casa **a su hermano**

Example of Majority Criterion



[change] [keep] [keep] [keep]

t_i: he missed his brother

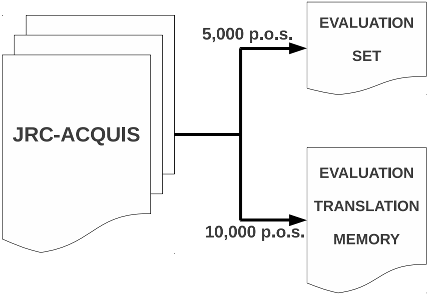
s_i: él echó de menos a su hermano

s': ella **echó de** casa **a su hermano**

Outline

1. Introduction
2. Related Work
3. Methodology
4. **Experiments and Results**
5. Conclusion
6. Current and future Work

Corpora


JRC-ACQUIS
5,000 p.o.s.
EVALUATION
SET
10,000 p.o.s.
EVALUATION
TRANSLATION
MEMORY

Evaluation Metrics

$$\text{Accuracy} = \frac{\text{correctly marked words}}{\text{marked words}}$$

$$\text{Coverage} = \frac{\text{marked words}}{\text{total words}}$$

Statistical Word Alignment

We use the GIZA++ (Och and Ney, 2003) free/open-source tool

- we obtain SL to TL alignment and a TL to SL alignment on the TM
- we experiment with three ways to combine the alignments:
  - union
  - intersection
  - grow-diag-final-and

Experimental Settings

We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++

Experimental Settings

We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++
- the use of both criteria defined to use the likelihood f_K (unanimity or majority)

Experimental Settings

We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++
- the use of both criteria defined to use the likelihood f_K (unanimity or majority)
- the use of alignment models trained on:

Experimental Settings

We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++
- the use of both criteria defined to use the likelihood f_K (unanimity or majority)
- the use of alignment models trained on:
  - the corpus to be aligned itself

Experimental Settings

We tried our approach comparing:

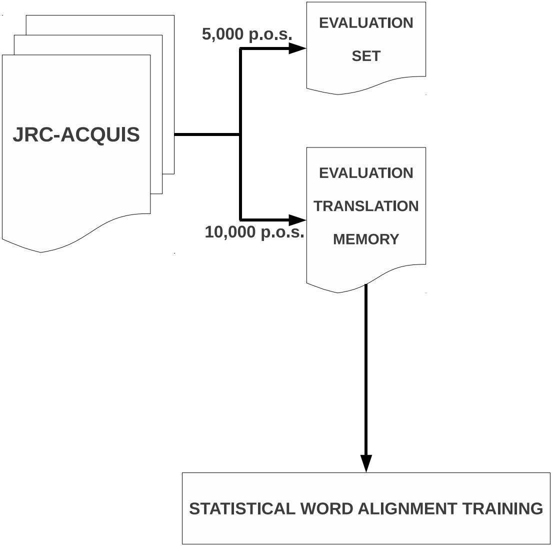
- the use of three different methods to combine the alignments generated with GIZA++
- the use of both criteria defined to use the likelihood f_K (unanimity or majority)
- the use of alignment models trained on:
  - the corpus to be aligned itself
  - a separate in-domain corpus

Experimental Settings

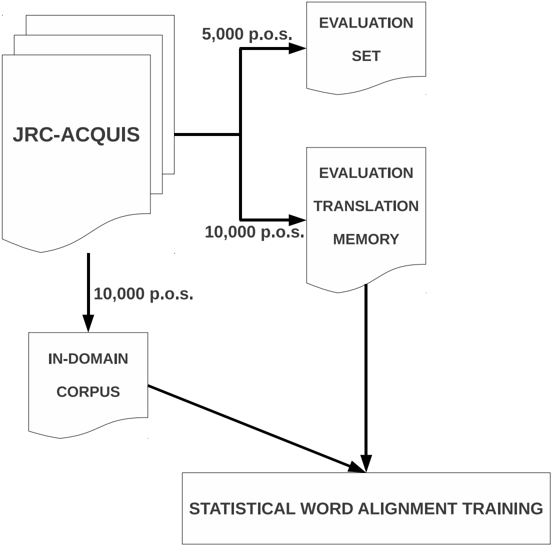
We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++
- the use of both criteria defined to use the likelihood f_K (unanimity or majority)
- the use of alignment models trained on:
  - the corpus to be aligned itself
  - a separate in-domain corpus
  - a separate out-of-domain corpus

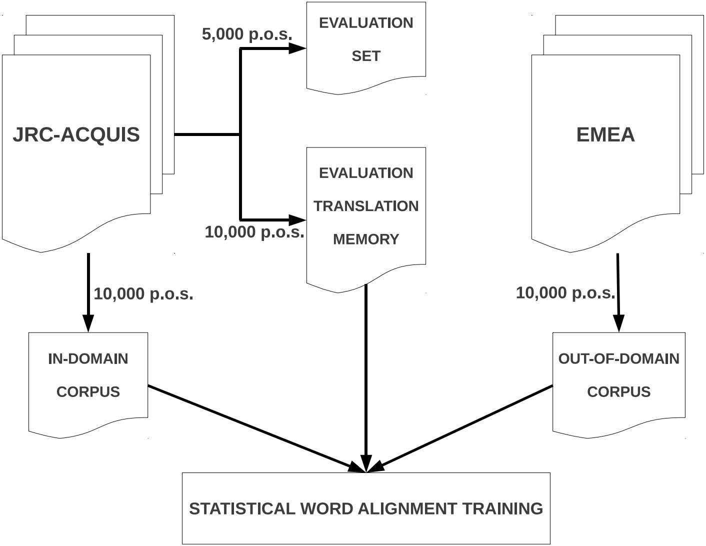
Corpora


JRC-ACQUIS
5,000 p.o.s.
EVALUATION SET
10,000 p.o.s.
EVALUATION TRANSLATION MEMORY
STATISTICAL WORD ALIGNMENT TRAINING

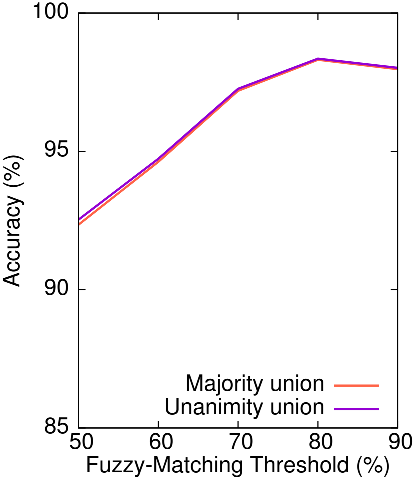
Corpora

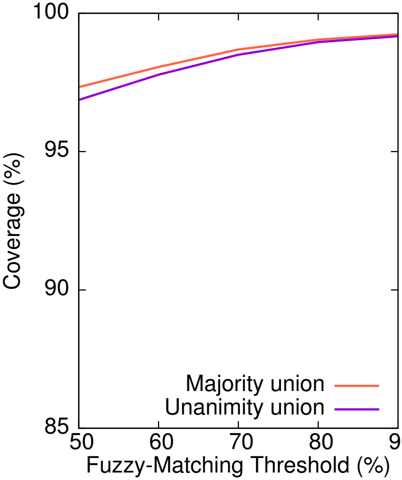

JRC-ACQUIS
5,000 p.o.s.
EVALUATION SET
10,000 p.o.s.
EVALUATION TRANSLATION MEMORY
10,000 p.o.s.
IN-DOMAIN CORPUS
STATISTICAL WORD ALIGNMENT TRAINING

Corpora

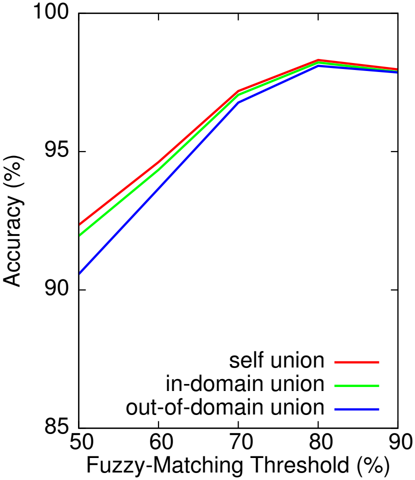

JRC-ACQUIS
5,000 p.o.s.
EVALUATION SET
10,000 p.o.s.
EVALUATION TRANSLATION MEMORY
EMEA
10,000 p.o.s.
IN-DOMAIN CORPUS
10,000 p.o.s.
OUT-OF-DOMAIN CORPUS
STATISTICAL WORD ALIGNMENT TRAINING

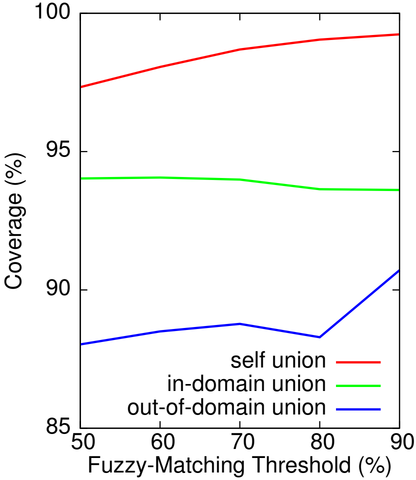
Results for the Majority/Unanimity Criteria





Results for the Different Alignment Models





Outline

1. Introduction
2. Related Work
3. Methodology
4. Experiments and Results
5. **Conclusion**
6. Current and future Work

Concluding Remarks

- new method to improve TM-based CAT tools
- predictability and high confidence of translators on fuzzy-match scores is kept
- accuracy over 94% for fuzzy match thresholds between 60% and 90%
- it is possible to reuse statistical alignment models from different corpora with a small loss in accuracy (but a larger loss in coverage)

Outline

1. Introduction
2. Related Work
3. Methodology
4. Experiments and Results
5. Conclusion
6. **Current and future Work**

Current and future Work

Current:

- surveying translators about the usefulness of target-side colouring (visit survey at `http://transducens.dlsi.ua.es/people/fsanchez/survey.html`)
- using MT to inform aligners and classifiers to colour target words in proposals *on the fly* (no need to train the aligner on a corpus)

Future:

- integration in the OmegaT free/open-source CAT system

License

HEEL ERG BEDANKT! MOLTES GRÀCIES!

Acknowledgements:

- Partially funded by the Spanish government through project TIN2009-14009-C02-01.
- Good ideas from Yanjun Ma, Andy Way and Harold Somers: thanks!