Enriching machine translation input using semantics-based fuzzy matches

Tom Vanallemeersch and Leen Sevens
Centre for Computational Linguistics
KU Leuven

CLIN 26, Amsterdam, December 2015
Overview

• Background:
  – Fuzzy matching
  – Semantic metrics
  – Combining matches and MT

• Semantic matches and MT
  – First experiment for English-Dutch
  – Adapting metric to Dutch-English

• Conclusions, future research

Research in SCATE and PICTO
Fuzzy matching

• CAT tools: retrieve similar sentences from TM
• Metrics:
  – String-based (e.g. Levenshtein)
  – Linguistics-based (e.g. comparison of syntax trees)
  – Originating from automated MT evaluation
  – Combined (Vanallemeersch and Vandeghinste 2015)

input: a total of 470 workers had accepted early retirement
TM match: many workers have taken early retirement
its translation: veel arbeiders gingen met vervroegd pensioen
Fuzzy matching

Some metrics are semantics-based:

• MT evaluation metrics: METEOR, MEANT
• Matching based on meaning rather than form
Semantic metrics

METEOR (Snover et al. 2006) contains several modules:

• Stemming
• Function words
• (Some languages) synonyms from WordNet, paraphrases

this resolution required the national parliament to present a report within a year
this resolution required the national parliament to present a document within a year
Semantic metrics

MEANT (Lo et al. 2012) performs comparison using SRL:
• Matching predicates
• Matching roles (role type should be identical)
• Lexical similarity (context vectors)

\[ A1 \text{ the relevant proposal} \] was \[ \text{pred presented} \] by \[ A0 \text{ the European Commission} \].

\[ A0 \text{ the Commission} \] \[ \text{pred presented} \] \[ A1 \text{ that proposal} \] \[ \text{AM-LOC there} \].

\[ \rightarrow \text{ Similar metric: SRL-Or (Giménez and Márquez 2010) } \]
Combining matches and MT

Determine translation of matching sentence parts:

• Detection of words that are aligned in TM sentence pair
• Word alignment (Koehn and Senellart 2010) → Levenshtein distance
• Subtree alignment (Zhechev and van Genabith 2010) → Combination of bilingual and monolingual alignment

input: a total of 470 workers had accepted early retirement
TM match: many workers have taken early retirement
its translation: veel arbeiders gingen op vervroegd pensioen
Combining matches and MT

Pretranslate using target part of links:

a total of 470 arbeiders had accepted early pensioen

Moses:

• XML markup of sentence parts
• Options:
  – Force Moses to use translation
  – Have it compete with other alternatives (LM decides)
Semantic matches and MT

Pretranslation using sentences with similar meaning:

- Part of research in SCATE project
- Language pair: English-Dutch
- Word alignment
- METEOR, MEANT
- Question: MT output improvement with respect to non-pretranslated input / Levenshtein?
Semantic matches and MT

Pretranslation using METEOR:

• Use of all English modules
• Matching part = sequence of identical/related words
• Translation of matching part = consistently word-aligned span

input: they were looking at a huge car
TM match: they will buy an enormous car
its translation: ze zullen een enorme auto kopen

input + markup: ze were looking at een enorme auto
Semantic matches and MT

Pretranslation using MEANT:

• Own implementation of the metric
  → Currently simplified: lexical overlap instead of similarity

• Matching parts:
  – Matching predicates
  – Identical words in matching roles with same type

input: [A0 they] were [pred looking] at [A1 a huge car]
TM match: [A0 they] [pred looked] at [A1 several cars] [AM-TMP today]
its translation: ze keken vandaag naar meerdere auto’s

input + markup: ze were keken at a huge auto’s
Semantic matches and MT

Experiment:

• Measure effect of semantic metrics and Levenshtein on MT output for Europarl v7

• Steps:
  – Create English-Dutch Moses model
  – Apply SRL (Johansson and Nugues 2008) to 70k subset of training sentences and 2k test sentences (from WMT shared task 2006)
  – Match test and training sentences (+ initial coarse filter); threshold 0.2
  – Detect translation-equivalent parts using GIZA++
  – Add markup to test sentences based on best match
  – Calculate BLEU, TER
Semantic matches and MT

Results:
• Scores similar to baseline
• Likely causes:
  – Match threshold too low (too much noise in markup)
  – Matching parts too small (often a single word)
  – Too little training sentences used for matching
  – Only lexical overlap (no similarity) in current MEANT implementation
  – Sparse word alignment
• Small matching parts have unpredictable outcome
Semantic matches and MT

Example of markup leading to positive result:

input: the convention did a grand job .
Moses output: het verdrag is een grote taak .

input + METEOR: the convention did a grand job .
best TM match: they have done a good , serious job .
its translation: zij hebben gedegen en serieus werk geleverd .

input + markup: the convention did a grand werk geleverd .
Moses output: de Conventie heeft een geweldige werk geleverd .
Semantic matches and MT

Example of synonymy with baseline output:

input:  ... devoted to reviewing the ...
Moses output:  ... gewijd aan de evaluatie van de ...

input + METEOR:  ... devoted to reviewing the ...
best TM match:  ... devoted to follow-up of the ...
its translation:  ... gewijd is aan de follow-up van de ...

input + markup:  ... gewijd is aan de follow-up the ...
Moses output:  ... gewijd is aan de follow-up van de ...
Semantic matches and MT

Example of MEANT match:

input: ... the European Council is due to adopt several decisions ...
Moses output: ... is de Europese Raad bevatten verschillende besluiten ...

input + MEANT: ... the European Council is due to adopt several decisions ...
best TM match: ... led the Council to adopt two important decisions ...
its translation: ... hebben ertoe geleid dat de Raad twee belangrijke besluiten heeft genomen ...

input + markup: ... de European Raad is due to adopt several besluiten ...
Moses output: ... is de Europese Raad bevatten verschillende besluiten ...

no alignment for adopt
Adapting metric to Dutch-English

PICTO project: METEOR + Dutch synonyms
• Purpose: text to pictograph conversion
• Lemmas from same synset, related lemma pairs:
  e.g. hyperonymy: kat – huisdier
• Based on Cornetto and Open Dutch WordNet (2014)

SCATE project: METEOR + Dutch paraphrases
• Purpose: fuzzy matching, inducing ASR probabilities, ...
• Derived from Moses phrase tables (Denkowksi 2014)
  e.g. house translates to both huis and woning, with some probability
Conclusions and future research

• Qualitative evaluation of MT output based on semantic matches shows some potential
• However, we need to control the effect of semantic metrics in a better way
• And need to enlarge/vary the TM for fuzzy matching
  → Full Europarl, “true” TMs like DGT and company TMs
• We will also work on Dutch as a source language
  → Adaptation of METEOR
Questions
References


