Choosing the best machine translation system to translate a sentence by using only source-language information

Felipe Sánchez-Martínez

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

fsanchez@dlsi.ua.es

15th Annual Conference of the
European Association for Machine Translation

May 30, 2011

Outline

Motivation /1

Multi-engine MT systems

- combine the output of N MT systems
  - alternatively they may first select a reduce set of translations $M < N$
- or select just one translation from the N computed ones

Drawbacks:

- N different translations must always be computed
- response time and amount of resources
- N needs to be kept to a minimum

Motivation /2

Goal

To select the MT system or subset of MT systems to use in advance, without translating and without access to the inner workings of the MT systems

Advantages:

- number of translations is drastically reduced
  $\Longrightarrow$ computing resources are saved
- focus on the combination of the best translations
- the number of MT systems N could be increased

System selection approach

The problem is faced as a classification approach that uses a set of source language (SL) features

- use of maximum entropy classifiers
- train a binary classifier per MT system
- use of parallel corpora and sentence-level MT evaluation metrics for training

Features obtained from the parse tree

Try to describe the sentence in terms of the complexity of its syntactic structure

- maximum number of child nodes
- mean number of child nodes
- number of internal nodes
- $p(t|w)$: likelihood of the parse tree given the words
- ...

System selection approach: SL features /2

Features related to the shift of the words and their fertilities

Try to describe the sentence in terms of the complexity of its words

shift : $\text{shift}(i) = \text{abs}(j - i)$
i: position of a SL word
j: position of the first TL word to which *i* is aligned

fertility : number of TL words to which a SL word is aligned

Several features. Number of words whose ...

- ... mean shift is above threshold Θ_1
- ... variance over the shift is above threshold Θ_2
- ... mean fertility is above threshold Θ_3
- ... variance over the fertility is above threshold Θ_4

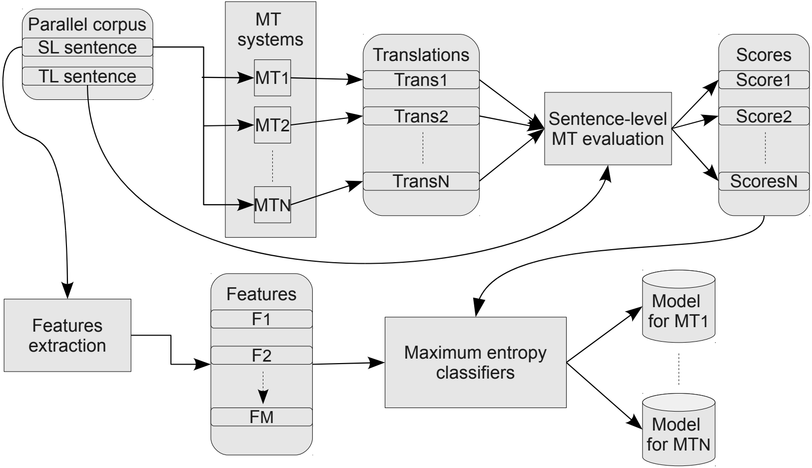
System selection approach: SL features /3

Other features

Try to discriminate between the rule-based MT systems and the corpus-based ones

- sentence length (in words)
- number of words not appearing in the corpora used to train the corpus-based systems
- likelihood of the sentence to translate as provided by a 5-gram language model trained on the corpora used to to train the corpus-based systems

System selection approach: training /1



System selection approach: training /2

Preprocessing

1. translate each SL sentence into the TL through all the MT systems
2. evaluate each translation against the reference translation in the training parallel corpus
3. determine the MT systems producing the best translation
   - several MT systems may produce the same translation, or several translations may be assigned the same score

System selection approach: training /3

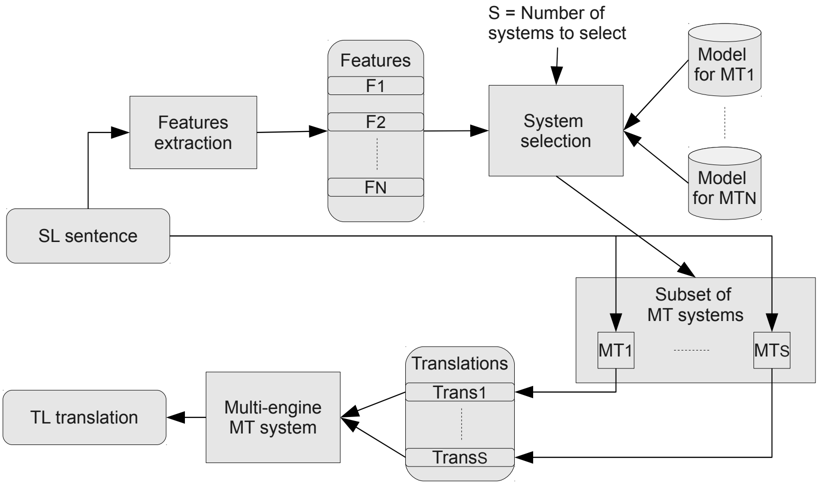
Training instances per MT

- one instance per parallel sentence in the training corpus
- if the MT is one of those producing the best translation(s) $\Longrightarrow$ that instance is classified as belonging to the class represented by that system

Training procedure

1. rank for each system all the features according to their chi-squared statistic with respect to the classes
2. train the different binary maximum entropy classifiers for the first F features in the ranking
3. determine the optimum value of F on a development corpus

System selection approach: selection /1



System selection approach: selection /2

System selection

1. compute the probability of each MT system being the best system to translate that sentence
2. select the subset of MT systems with the highest probabilities
   - in the experiments we select only one system, the one with the highest probability

Experimental settings /1

Translation of English and French texts into Spanish

MT systems

- Apertium (Forcada et al., 2011) rule-based MT
- Moses (Koehn et al., 2007) phrase-based statistical MT
- Moses hierarchical phrase-based statistical MT (Chiang, 2007)
- Cunei (Phillips and Brown, 2009) hybrid example-based–statistical MT
- Yahoo! Babelfish (systran) rule-based MT

Experimental settings /2

Corpora

- corpus-based systems trained on the Europarl and News Commentary corpora released for WMT10
- training, development and test corpora: UN corpus released for WMT10

Pair	Corpus	Num. sent.	Num. words
en-es	Train	98,480	en: 2,996,310; es: 3,420,636
	Dev	1,984	en: 49,003; es: 57,162
	Test	1,985	en: 55,168; es: 65,396
fr-es	Train	99,022	fr: 3,513,404; es: 3,449,999
	Dev	1,987	fr: 60,352; es: 59,551
	Test	1,982	fr: 64,392; es: 64,440

Experimental settings /3

Other resources

- Berkeley Parser (Petrov et al., 2006)
- IRSTLM language modelling toolkit (Federico et al., 2008)
  - 5-gram language model trained on the SL Europarl and News Commentary corpora
- Asiya evaluation toolkit (Giménez and Màrquez, 2010)
  - Evaluation metrics: BLEU, PER, TER, METEOR
- WEKA machine learning toolkit (Witten and Frank, 2005)

Results and discussion /1

Pair	Configuration	BLEU	TER	METEOR
`en-es`	Best system	0.3481	0.4851	0.2745
	System selection	0.3529	0.4838	0.2762
	Oracle	0.3905	0.4409	0.2965
`fr-es`	Best system	0.3146	0.5880	0.2281
	System selection	0.3192	0.5861	0.2286
	Oracle	0.3467	0.5548	0.2389

Oracle translation: for each sentence, the translation with the highest score (at the sentence level) is chosen

Best system: System performing best at the document level

Results and discussion /2

- 95% confidence intervals computed by 1,000 iterations of bootstrap resampling show a large overlapping between "System selection" and "Best system"
- No overlapping between "System selection" and "Oracle"
- Results are statistically significant according to pair bootstrap resampling (except for `fr-es` and METEOR)

Percentage of times each systems is chosen when translating the test corpora

Pair	Measure	PMOS	HMOS	CUNE	APER	SYST
en-es	BLEU	32.9%	51.1%	2.6%	0.1%	13.3%
	TER	53.6%	36.0%	5.5%	0.0%	4.9%
	METEOR	28.8%	18.5%	41.8%	0.0%	10.9%
fr-es	BLEU	0.2%	42.5%	38.1%	0.0%	19.2%
	TER	0.2%	36.7%	53.7%	0.0%	9.4%
	METEOR	0.0%	26.6%	63.2%	0.0%	10.2%

Results and discussion /4

Inspection of the first 500 sentences in the `en-es` test corpus

- most of the times the MT systems produce translations of similar quality
- manual ranking of the automatic translations without access to the reference translations

Configuration	BLEU
Best system	0.3926
Manual selection	0.3928

Possible reason

- the three corpus-based systems were trained on the same parallel corpora

Further experiments: work in progress

Trying with additional corpus-based systems trained on different corpora $\Longrightarrow$ 12 systems in total

- EMEA (medical domain)
- JRC-Acquis (legal domain)
- OpenSubtitles (open domain)

Preliminary evaluation results

in-domain The improvement with respect to the MT performing best at the document level is larger

out-of-domain No improvement is obtained as compared to the MT performing best at the document level

Concluding remarks and future work

Remarks

- Novel approach aimed to select the subset of MT systems to use by multi-engine MT systems in advance, without translating
- Only SL information is used
- Preliminary experiments on two language pairs show a small improvement when evaluated with in-domain data

Future work

- try other classification approaches
- think of additional features
- select a subset of systems (instead of just one) and combine their translations using MANY (Barrault, 2010)

Choosing the best machine translation system to translate a sentence by using only source-language information

Felipe Sánchez-Martínez

Work funded by the EAMT through its 2010 sponsorship of activities program

Thank you very much for your attention!
Dank u zeer voor uw aandacht!

May 30, 2011