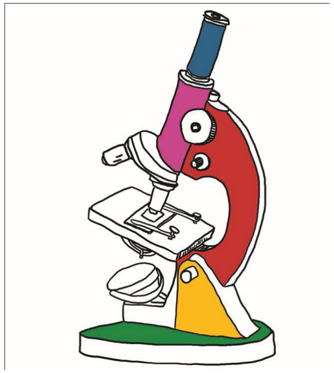
The concept and feasibility of modern statistical machine translation

Domain Adaptation



Maxim Khalilov
TAUS Labs
Amsterdam

Marta R. Costa-jussà
Barcelona Media
Barcelona

RuSSIR 2012
August 5-10

Motivation

Outline

- Motivation
- Statistical approaches
  - Translation/Language model interpolation (Koehn and Schroeder, 2007) and Mixture models (Banerjee, 2011)
  - Training data selection (Lü and Liu, 2007)
  - Weighting and combination of multiple translation resources (Rogati, 2009)
  - Translation edit rate (Henriquez, 2011)

Outline

- Motivation
- Statistical approaches
  - Translation/language model interpolation (Koehn and Schroeder, 2007) and Mixture models (Banerjee, 2011)
  - Training data selection (Lü and Liu, 2007)
  - Weighting and combination of multiple translation resources (Rogati, 2009)
  - Translation edit rate (Henriquez, 2011)

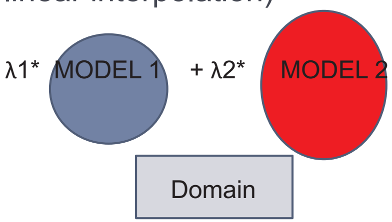
Corpora from different domains allow for straightforward combination alternatives

- Concatenation of in-domain and out-domain corpus
- In-domain language model: LM is trained only with in-domain corpus

(Koehn and Schroeder, 2007)

Model interpolation allows to use all training data but include a preference for the in-domain jargon

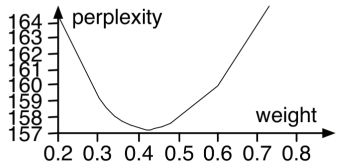
- Interpolated language model (linear interpolation)
- Two language models (log linear interpolation)
- Two translation models (log linear interpolation)



Preference for the in-domain jargon is achieved by giving more weight to the in-domain data

Interpolation of language models is optimized minimizing perplexity

- Expectation-Maximization optimization algorithm



Interpolation can be done directly in the decoder

- Two language models are included as two separate features, whose weights are set with minimum error rate training
- Two translation models are introduced taking advantage of the Moses decoder's factored translation model framework.
  - It is possible to use multiple alternative decoding paths (Birch, 2007)

Using two translation and language models is the best alternative

DATA

Data set	Fr-En
Europarl	1,257,419
News	42,884
Development	2,000
Test (NEWS)	2,007

RESULTS

Method	%BLEU
Combined training data	26.69
In-domain language model	27.46
Interpolated language model	27.12
Two language models	**27.30**
Two translation models	**27.64**

Further experiments with interpolation: Mixture Models

- Interpolation of translation and language models can be done linearly or log linearly.

(Banerjee et al, 2011)

Linear interpolation is the best alternative

DATA

Data set	Fr-En
Symantex TM	567,641
Europarl	414,667
Development Set	500
Test Set	612
English forum	1,069,464

RESULTS

TM	LM	BLEU
TM	TM+forum	36.42
TM+EP	Conc	36.81
TM+EP	Linmix	36.92
TM+EP	Logmix	36.74
Linmix	Conc	36.56
Linmix	**Linmix**	**37.10**
Linmix	Logmix	36.74
Logmix	Conc	34.88
Logmix	Linmix	36.52
logmix	Logmix	36.39

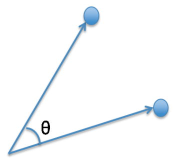
Outline

- Motivation
- Statistical approaches
  - Translation/Language model interpolation (Koehn and Schroeder, 2007) and Mixture models (Banerjee, 2011)
  - Training data selection (Lü and Liu, 2007)
  - Weighting and combination of multiple translation resources (Rogati, 2009)
  - Translation edit rate (Henriquez, 2011)

Training data selection using information retrieval techniques

- Select train sentences similar to the test using cosine similarity and tf-idf.
  - Cosine similarity

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



(Lü and Liu, 2007)

Term frequency- invers document frequency

- Term frequency (tf): the more a term appears in a sentence, the more relevant it is for that sentence
- Inverse document frequency (idf): the less a term appears in the other sentences, the more relevant it is.

Chinese-English results

System	Distinct pairs	Size training	BLEU
Baseline	600000	2.41G	23.63
Top100	91804	0.43G	23.06
Top200	150619	0.73G	23.60
Top500	261003	1.28G	24.15
Top1000	357337	1.74G	**24.63**
Top2000	445890	2.13G	23.51

Outline

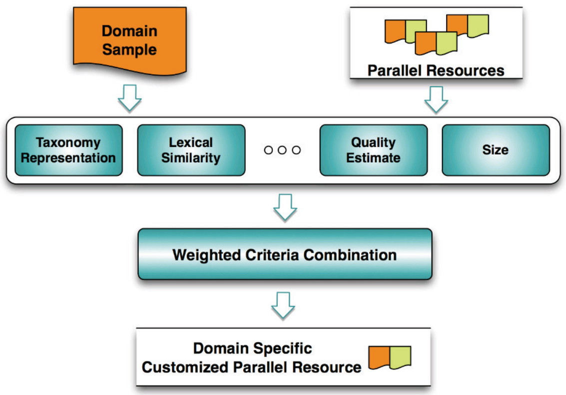
- Motivation
- Statistical approaches
  - Translation/Language model interpolation (Koehn and Schroeder, 2007) and Mixture models (Banerjee, 2011)
  - Training data selection (Lü and Liu, 2007)
  - Weighting and combination of multiple translation resources (Rogati, 2009)
  - Translation edit rate (Henriquez, 2011)

Domain adaptation gains interest when more resources are available

- It is interesting a flexible framework that given several parallel resources and a domain sample, produces a customized domain adapted parallel resource

(Rogati, 2009)

Adaptation Framework Overview



Redundancy

- A large quantity of redundant corpus diminishes utility regardless of how closely it matches the domain.
- Redundancy is measured by examining the lexical level similarity between the previously selected parallel sentences and a new candidate.

Lexical Similarity is a crucial domain match criterion

- Cosine similarity between two sets (p,r):

$$\cos(p,r) = \frac{\sum_w p(w)r(w)}{\sqrt{\sum_w p(w)^2}\sqrt{\sum_w r(w)^2}}$$

- Binary version of Jaccard's coefficient measures the overlap between two sets (p,r):

$$Jaac_b(p,r) = \frac{|pIr|}{|pYr|}$$

More similarity measures…

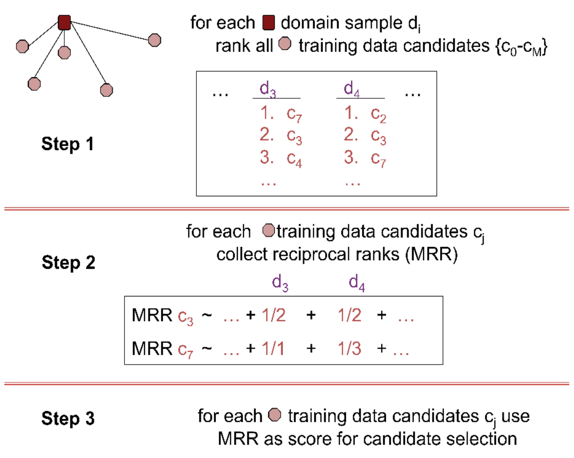
- Kullback-Leibler divergence

$$KL(p,r) = \sum_{w} p(w) log \frac{p(w)}{r(w)}$$

- Language model perplexity, given n sentences in the domain sample, we can calculate the Perp of the LM trained on the parallel corpus p

$$Perp = 2^{-\frac{1}{|p|}\sum_{i=1}^{n} logP\,(S_i)}$$

Similarity aggregation using the Mean Reciprocal Rank



Translation quality estimation

- Length/Ratio Variance: the ratio of the number of words in the original texts vs. the translated text

$$\sigma^2 = \frac{\sum_{i=1\ldots|C|}(\lambda_i - \mu_\lambda)^2}{|C|}$$

λ_i, the ratio of the lengths (in words of the i-th sentences in each half of C

μ_X, the mean of λ_i in C

$|C|$, the size of the collection in sentences or documents

- Bootstrap and evaluation:
  - Half of the parallel corpus is translated using another parallel corpus
  - BLEU to each sentence

More TQE …

- Translation probabilities stability: how term-to-term translation probabilites change when a random selection of docs is eliminated from training

$$\frac{\sum_{i=1..K}(\delta_i - \mu_\delta)^2}{K}$$

$$\delta_k = \frac{\sum_{i=1..|V_e|\,j=1..|V_f|}(p(e_i|f_i) - p_k(e_i|f_j))}{|V_e||V_f|}$$

K is the number of folds/turns in eliminating documents

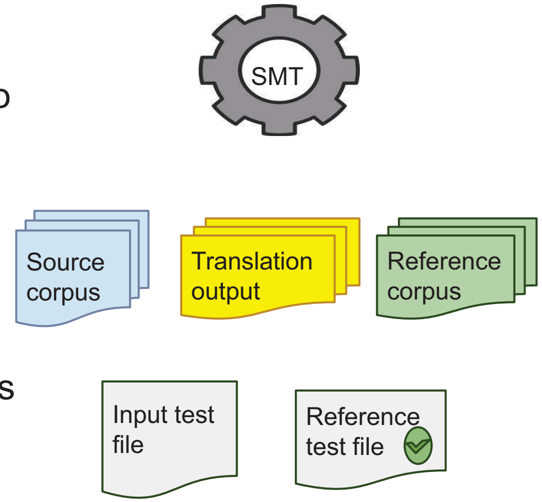
Size can be used as a thresholding measure

- Smaller sentences add little additional information
- Overly-long sentences lead to less sharp co-ocurrence probabilities used for the translation model

Domain adaptation can be done by means of translation edit rate

- A baseline system to adapt
- A derived corpus to adapt the system
- A parallel test corpus

(Henríquez et al, 2011)

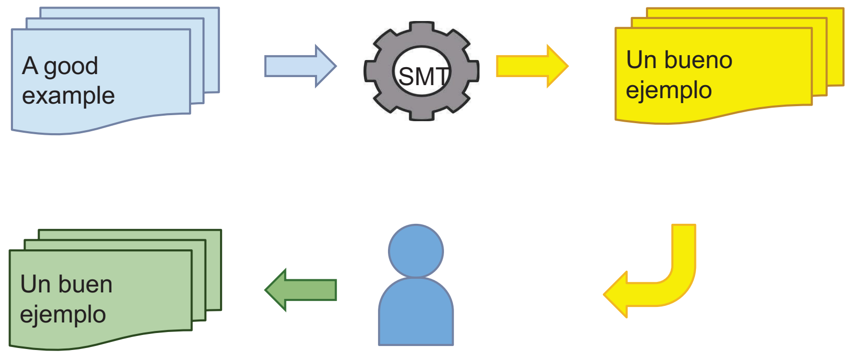


Outline

- Motivation
- Statistical approaches
  - Translation/Language model interpolation (Koehn and Schroeder, 2007) and Mixture models (Banerjee, 2011)
  - Training data selection (Lü and Liu, 2007)
  - Weighting and combination of multiple translation resources (Rogati, 2009)
  - Translation edit rate (Henriquez, 2011)

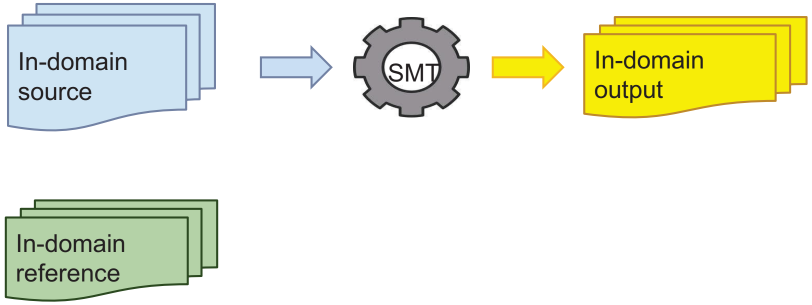
Origin of the derived corpus

- User's feedback



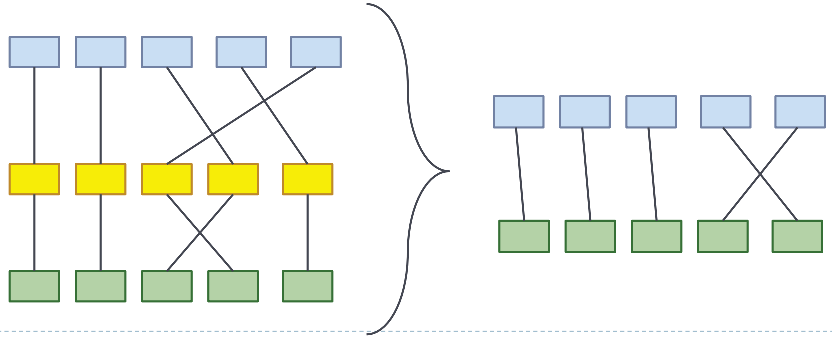
Origin of the derived corpus

- In-domain parallel corpus



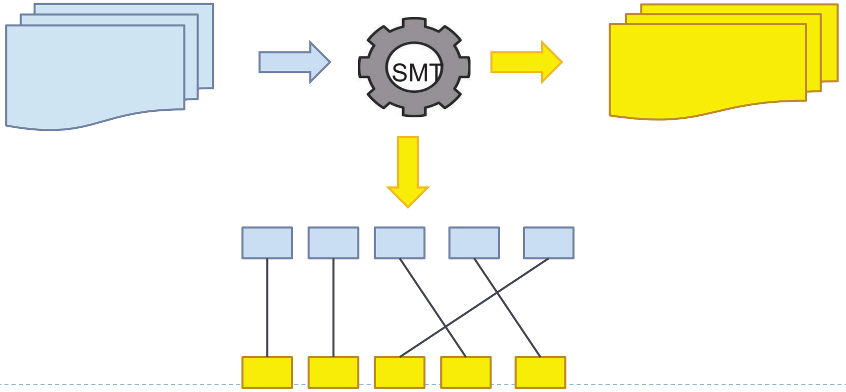
Objective

- Use the translation output as a pivot to align the input with the reference and extract new translation units



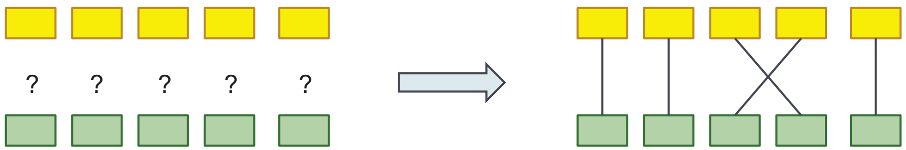
First step

- Links between input and translation output: provided by the SMT system



Second step

- Compare translation output with reference to automatically detect changes

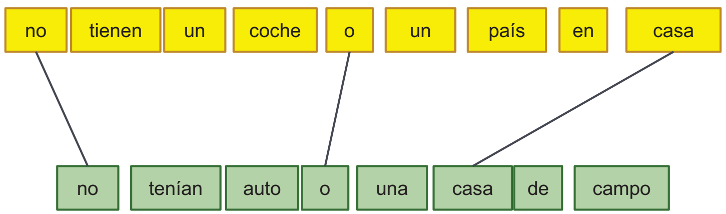


Second step: How-to

- First we detect and link the identical words that appear in both sentences using Translation Edit Rate (TER)
- Then we compare the remaining non-linked words using a similiarity measure and set the links with a greedy approach

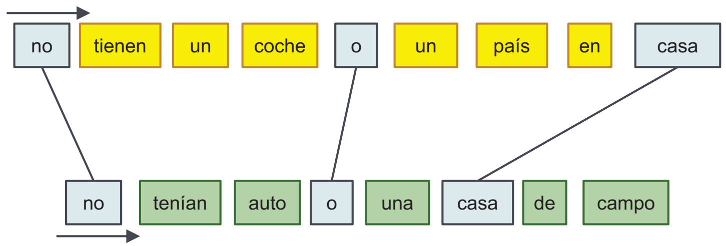
Detect identical words using TER

- TER computes the minimum number of edits (insertion, deletions and replacement) needed to change a sentence into another, allowing phrase shifting



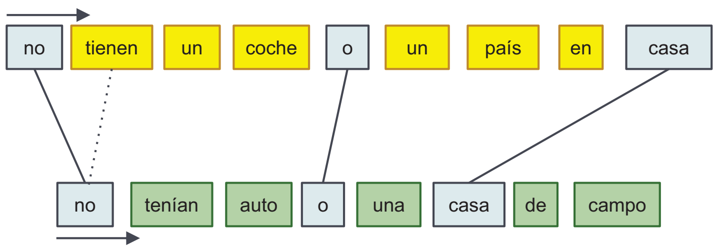
A similarity function for the remaining words

- Looping from left to right we iterate over all non-linked output words and all reference words, computing the similarity between them



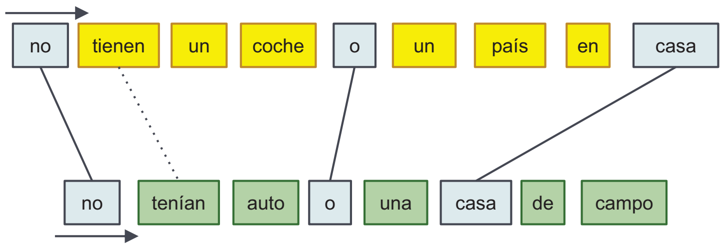
A similarity function for the remaining words

- The similarity function consider 6 different features to measure similarity



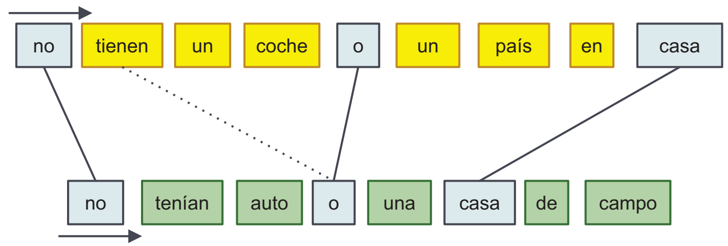
A similarity function for the remaining words

- Two features to measure the lexical relationship, considering the source words as origin of both



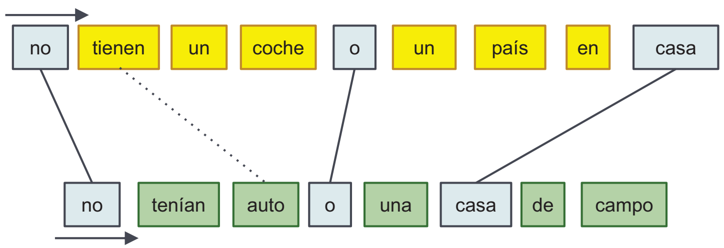
A similarity function for the remaining words

- One feature that penalizes the similarity if the reference word is already linked with an output word which is far from the current word



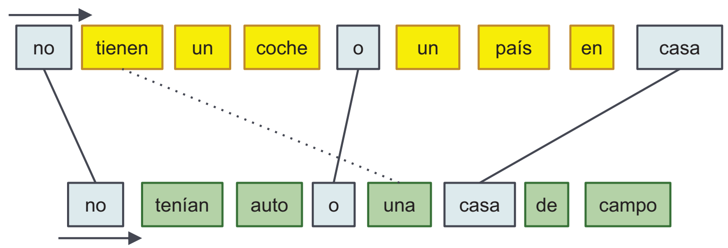
A similarity function for the remaining words

- One feature to check if the words are identical



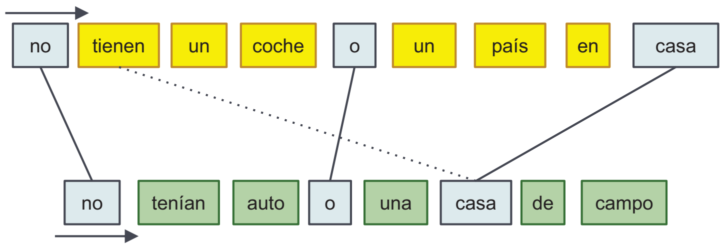
A similarity function for the remaining words

- One feature that consider if the previous output word is linked with the previous or next reference word



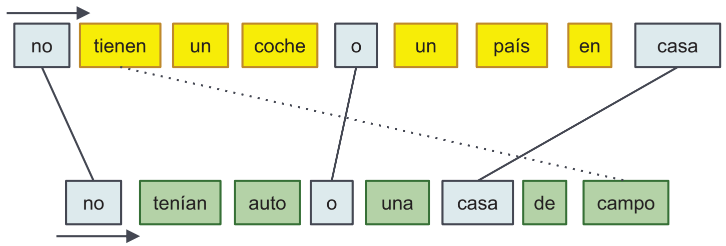
A similarity function for the remaining words

- One feature that consider if the next output word is linked with the previous or next reference word



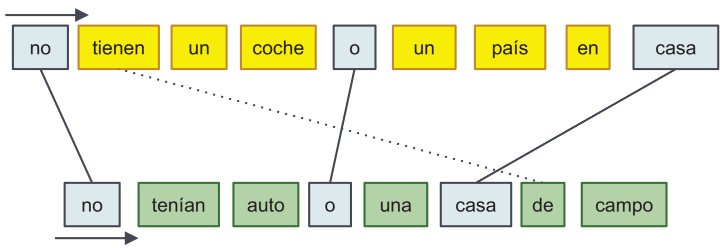
A similarity function for the remaining words

- At the end the decision is taken using a greedy approach



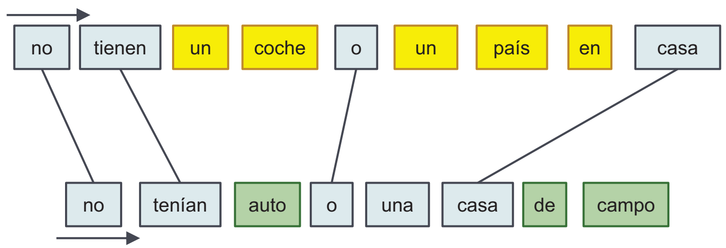
A similarity function for the remaining words

- After all features are computed for a pair of words, they are linearly combined to obtain the final similarity value



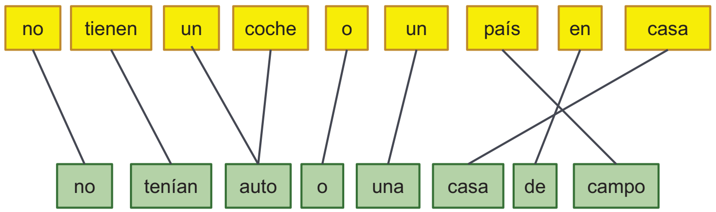
A similarity function for the remaining words

- Therefore the final link is assigned to the pair that obtained the maximum similarity value



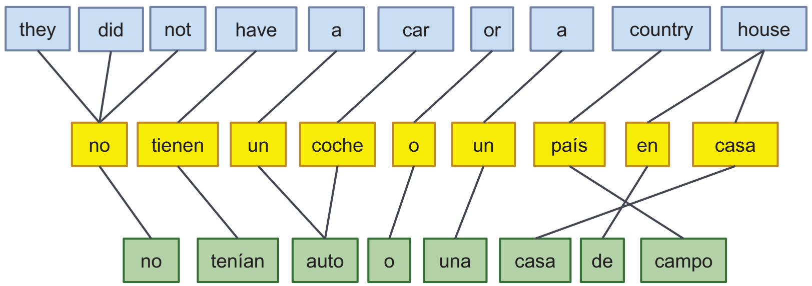
A similarity function for the remaining words

- At the end, all output words will be linked with one reference word



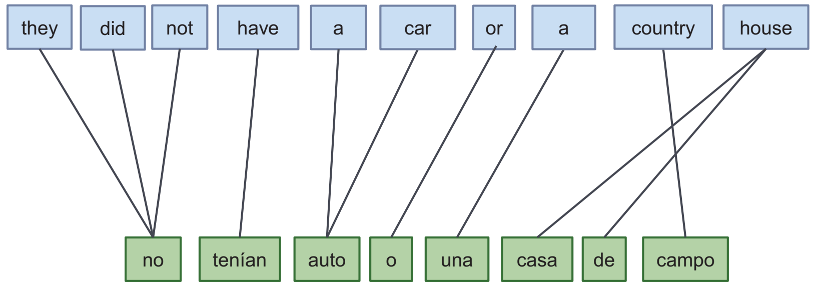
Third step

- With all links computed with use the translation output as pivot to obtain a word alignment



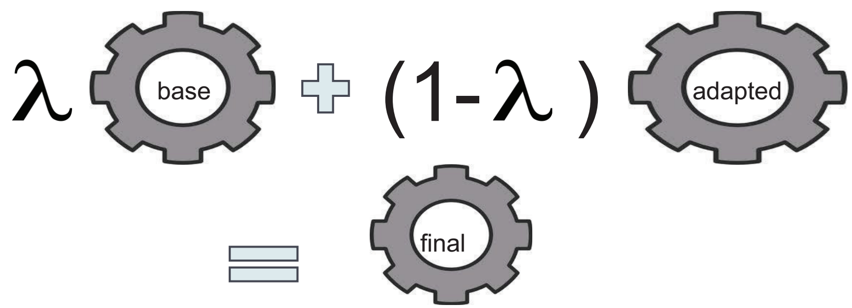
Fourth step

- Once we've obtained the alignment we extract all phrases and build an adapted model using the standard tools



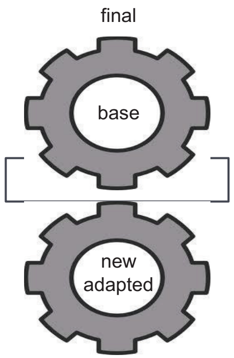
Building the final translation model

- The final TM model is computed with a linear combination
- The combination will add new phrases and adapt the remaining ones accordingly



Building the final reordering model

- The final reordering model is the baseline model augmented with the new phrases found in the adapted model



Next: recent advances in SMT

- Did you understand the basic theory??? Let's go a little bit further…
  - Factored translation models
  - N-gram-based translation models
  - Hiero
  - Syntax-based translation systems