

Exploiting Lexical Information and Discriminative Alignment Training in Statistical Machine Translation

Patrik Lambert

(supervisor: Rafael E. Banchs)

(tutor: Núria Castell)

TALP Research Center
Jordi Girona Salgado, 1-3
08034 Barcelona, Spain

UPC, April 2008

- 1 Introduction
- 2 N-gram-based Machine Translation
- 3 Word Alignment (Summary)
- 4 Multi-word Expression Grouping (Summary)
- 5 Parameter Optimisation Improvements (Summary)
- 6 Alignment Minimum-Translation-Error Training
- 7 Conclusions and Further Work

The Statistical Approach to Machine Translation

- Noisy channel model [Brown et al. 93]:

$$\hat{\mathbf{t}} = \arg \max_{\mathbf{t}} Pr(\mathbf{t} | \mathbf{s}) = \arg \max_{\mathbf{t}} Pr(\mathbf{t}) Pr(\mathbf{s} | \mathbf{t})$$

- Word-based translation model
 - Hidden alignment introduced in the translation model
 - Alignment: describes correspondences between words
- Generalisation to log-linear combination of feature functions [Och and Ney 02]

$$\hat{\mathbf{t}} = \arg \max_{\mathbf{t}} \sum_m \lambda_m h_m(\mathbf{s}, \mathbf{t})$$

- Units: words \rightarrow phrases
- Phrase-based SMT [Zens et al. 02, Koehn et al. 03], N-gram MT:
 - No alignment parameters in translation model (TM)
 - Units extracted from Viterbi alignments computed in previous stage

Main Objectives

- Study on Word Alignment Evaluation, Influence of Reference Corpus

Main Objectives

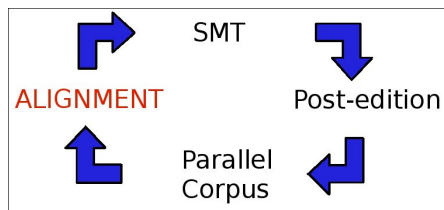
- Study on Word Alignment Evaluation, Influence of Reference Corpus
 - SMT systems: word-based → phrase-based but alignment **word-based**
 - non-compositional multi-word expressions (MWEs) can't be aligned by word-based models
 - example: *fire engine - camión de bomberos*
- ⇒ Extract MWEs automatically, group them as a unique token before alignment training.

Main Objectives

- SMT model weight optimisation process creates two types of problems:
 - Accuracy problem: translation biased towards the metric used
⇒ SMT parameter adjustment based on reliable criterion: QUEEN measure of Qarla Framework (IQmt toolkit [Giménez and Amigó 06])
 - works with a combination of metrics
 - gives probability that for every metric, the automatic translation is closer to a reference than 2 references to each other
 - Local optima ⇒ experimental error introduced
⇒ better algorithm to reduce variability of results (SPSA algorithm, compared Downhill Simplex method)

Main Objectives

- Alignment and SMT training are two separated steps: alignment not performed as function of SMT system
- SMT not easy to customize. Good framework to alleviate this drawback:



Bottleneck: alignment stage (long time required to align a small corpus increment)

⇒ **flexible** alignment framework adequately **adapted** to the machine translation task

- 1 Introduction
- 2 N-gram-based Machine Translation**
- 3 Word Alignment (Summary)
- 4 Multi-word Expression Grouping (Summary)
- 5 Parameter Optimisation Improvements (Summary)
- 6 Alignment Minimum-Translation-Error Training
- 7 Conclusions and Further Work

The translation model

- N-gram language model of bilingual units (tuples)

$$p(S, T) = \prod_{k=1} p((\tilde{s}, \tilde{t})_k | (\tilde{s}, \tilde{t})_{k-N+1}, \dots, (\tilde{s}, \tilde{t})_{k-1})$$

- Tuples extracted from word alignment; set of smallest phrase pairs
 - which form a monotonous segmentation of each sentence pair
 - such that no word in a tuple is aligned to words outside of it



Tuples:

1.- NULL : we

2.- quisieramos : would like

3.- lograr : to achieve

4.- traducciones perfectas : perfect translations

- From the translation model to the translation system: more features+optimisation tool

Selected Publications: N-gram Machine Translation

- José B. Mariño, Rafael Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, Marta R. Costa-jussà. **N-gram-based Machine Translation**. Computational Linguistics, 2006

- 1 Introduction
- 2 N-gram-based Machine Translation
- 3 Word Alignment (Summary)**
- 4 Multi-word Expression Grouping (Summary)
- 5 Parameter Optimisation Improvements (Summary)
- 6 Alignment Minimum-Translation-Error Training
- 7 Conclusions and Further Work

Alignment Task

- Popular alignment methods:
 - generative models trained with EM algorithm [Brown *et al.* 93]
GIZA++ tool ; combine source-target and target-source alignments
 - based on association measures (interpret cooccurrence frequency data)
 - supervised methods (with lexical models trained in unsupervised way)
discriminative systems (log-linear models) [Fraser and Marcu 05, Liu *et al.* 05, Ittycheriah and Roukos 05, Moore 05]
- Alignment Evaluation:
 - Alignments are compared to a manually aligned reference corpus
 - precision (P): proportion of computed links present in the reference
 - recall (R): proportion of reference links that were computed
 - Alignment Error Rate (AER) [Och and Ney 00]: generalises $1 - F$ with unambiguous links (S or Sure) and ambiguous links (P or Possible)

Alignment Task

- Popular alignment methods:
 - generative models trained with EM algorithm [Brown *et al.* 93]
GIZA++ tool ; combine source-target and target-source alignments
 - based on association measures (interpret cooccurrence frequency data)
 - supervised methods (with lexical models trained in unsupervised way)
discriminative systems (log-linear models) [Fraser and Marcu 05, Liu *et al.* 05, Ittycheriah and Roukos 05, Moore 05]
- Alignment Evaluation:
 - Alignments are compared to a manually aligned reference corpus
 - precision (P): proportion of computed links present in the reference
 - recall (R): proportion of reference links that were computed
 - Alignment Error Rate (AER) [Och and Ney 00]: generalises $1 - F$ with unambiguous links (S or Sure) and ambiguous links (P or Possible)
- AER sensitive to S/P links ratio in the reference corpus. In particular
 - if $|\{\text{P links}\}| \gg |\{\text{S links}\}|$, high precision alignments are favoured
 - if $|\{\text{S links}\}| \gg |\{\text{P links}\}|$, high recall alignments are favoured

Selected Publications: Language Resources and Evaluation

- Victoria Arranz, Núria Castell, Josep M. Crego, Jesús Giménez, Adrià de Gispert and Patrik Lambert. **Bilingual Connections for Trilingual Corpora: An XML Approach**. LREC 2004
- Patrik Lambert, Adrià de Gispert, Rafael Banchs and José B. Mariño. **Guidelines for Word Alignment Evaluation and Manual Alignment**. Language Resources and Evaluation, 2005
- Rafael E. Banchs, Josep M. Crego, Patrik Lambert, José B. Mariño. **A Feasibility Study For Chinese-Spanish Statistical Machine Translation**. ISCSLP 2006

Correlation with Machine Translation Quality

Collaboration with de Gispert (UPC), RWTH, ITC-IRST

- corpus morphology reduction:
 - *baseline**: true case, without classes, Giza 1⁵ HMM⁵ 3³ 4³
 - *baseline*: lower case, 50 classes, Giza 1⁴ HMM⁵ 4⁴
 - *full verbs*: group the whole verbal form (including pronouns, auxiliary verbs and head verb) into the lemma of the head verb
 - *stems*: replace word form by stem (suffix or derivation is removed)
- original forms restored after alignment: changes only affect alignment

	AER	Eng→Spa	Spa→Eng
		BLEU	BLEU
baseline*	20.9	0.479	0.552
baseline	17.6	0.480	0.553
full verbs	17.0	0.479	0.551
stems	16.7	0.479	0.555
verbs + stems	16.4	0.478	0.552

Correlation with Machine Translation Quality

Collaboration with de Gispert (UPC), RWTH, ITC-IRST

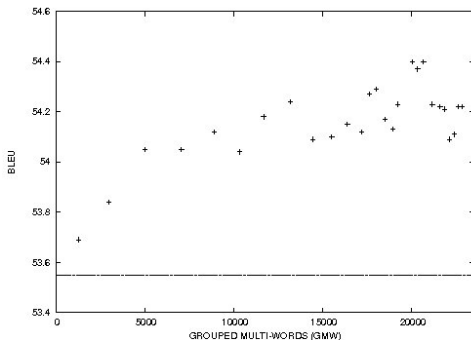
- corpus morphology reduction:
 - *baseline**: true case, without classes, Giza 1⁵ HMM⁵ 3³ 4³
 - *baseline*: lower case, 50 classes, Giza 1⁴ HMM⁵ 4⁴
 - *full verbs*: group the whole verbal form (including pronouns, auxiliary verbs and head verb) into the lemma of the head verb
 - *stems*: replace word form by stem (suffix or derivation is removed)
- original forms restored after alignment: changes only affect alignment

	AER	Eng→Spa	Spa→Eng
		BLEU	BLEU
baseline*	20.9	0.479	0.552
baseline	17.6	0.480	0.553
full verbs	17.0	0.479	0.551
stems	16.7	0.479	0.555
verbs + stems	16.4	0.478	0.552

- 1 Introduction
- 2 N-gram-based Machine Translation
- 3 Word Alignment (Summary)
- 4 Multi-word Expression Grouping (Summary)**
- 5 Parameter Optimisation Improvements (Summary)
- 6 Alignment Minimum-Translation-Error Training
- 7 Conclusions and Further Work

Multi-word Expressions Grouping

- Method to extract automatically bilingual MWEs
- 2 experiments suggest that SMT improved by grouping MWEs before alignment (Giza ++):
- Verbmobil corpus:



Multi-word Expressions Grouping

- EPPS corpus:
 - extraction method probably too noisy: not only non-compositional MWE grouped
 - not quantitative improvement overall
 - detailed error analysis: when MWE grouped were non-compositional, grouping helped for their correct translation

Selected Publications: Linguistic Classification and Multi-word Expressions

- Adrià de Gispert, Deepa Gupta, Maja Popovic, Patrik Lambert, José B. Mariño, Marcello Federico, Hermann Ney and Rafael Banchs. **Improving Statistical Word Alignments with Morpho-syntactic Transformations.** FinTAL 2006
- Patrik Lambert and Núria Castell. **Alignment of parallel corpora exploiting asymmetrically aligned phrases.** LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora
- Patrik Lambert and Rafael Banchs. **Data Inferred Multi-word Expressions for Statistical Machine Translation.** Machine Translation Summit X (2005)
- Patrik Lambert and Rafael Banchs. **Grouping Multi-word Expressions According to Part-Of-Speech in Statistical Machine Translation.** EACL 2006 Workshop on Multi-Word-Expressions in a Multilingual Context

- 1 Introduction
- 2 N-gram-based Machine Translation
- 3 Word Alignment (Summary)
- 4 Multi-word Expression Grouping (Summary)
- 5 Parameter Optimisation Improvements (Summary)**
- 6 Alignment Minimum-Translation-Error Training
- 7 Conclusions and Further Work

SMT System Tuning according to QUEEN measure

Collaboration with UPC-LSI and UNED (Madrid)

- Compared SMT model weight optimisation as a function of BLEU and QUEEN
 - QUEEN: probability that, for every metric, translation is closer to a reference than two other references to each other
 - according to manual evaluation, higher quality translations for system tuned as function of QUEEN than BLEU
 - example: negative sentence → all references contain “not”

Source	Creo que no se puede pensar que lo que gana una institución lo pierde la otra .
QUEEN tuning	I believe that it is not possible to think that what wins a institution what loses the other .
BLEU tuning	I believe that it is possible to think that what wins a institution loses the other .

Simultaneous Perturbation Stochastic Approximation

- The SPSA method [J. Spall, 1992] is based on a **gradient approximation** which requires only **two evaluations** of the objective function, regardless of the dimension of the optimisation problem.
- SPSA procedure is in the general recursive stochastic approximation form:

$$\hat{\lambda}_{k+1} = \hat{\lambda}_k - \mathbf{a}_k \hat{\mathbf{g}}_k(\hat{\lambda}_k)$$

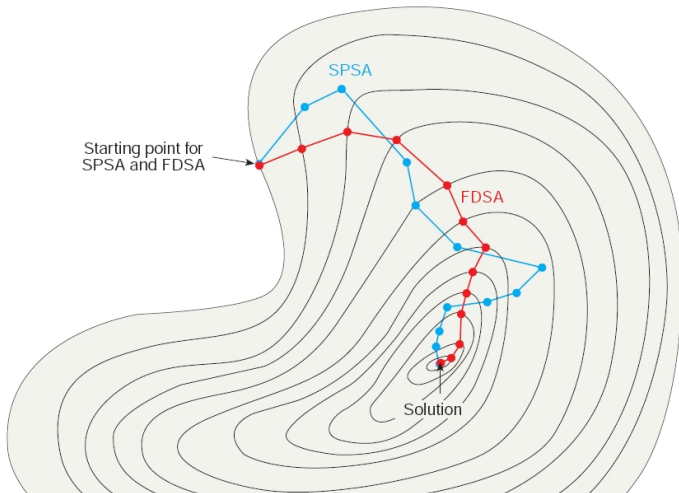
$\hat{\mathbf{g}}_k(\hat{\lambda}_k)$: estimate of the gradient $\mathbf{g}(\lambda) \equiv \partial E / \partial \lambda$ at iterate k

Simultaneous Perturbation Stochastic Approximation

- The SPSA method [J. Spall, 1992] is based on a **gradient approximation** which requires only **two evaluations** of the objective function, regardless of the dimension of the optimisation problem.
- SPSA procedure is in the general recursive stochastic approximation form:

$$\hat{\lambda}_{k+1} = \hat{\lambda}_k - \mathbf{a}_k \hat{\mathbf{g}}_k(\hat{\lambda}_k)$$

$\hat{\mathbf{g}}_k(\hat{\lambda}_k)$: estimate of the gradient $\mathbf{g}(\lambda) \equiv \partial E / \partial \lambda$ at iterate k



The simultaneous approximation causes deviations of the search path. These deviations are averaged out in reaching a solution.

Comparing SPSA to Downhill Simplex

- Raised issue of experimental error caused by minimum error training strategy
 - adapted SPSA algorithm to SMT system tuning problem
 - variability of results reduced w.r.t. Downhill Simplex method

Selected Publications: Parameter Optimisation

- Patrik Lambert and Rafael E. Banchs. **Tuning Machine Translation Parameters with SPSA**. IWSLT
- Patrik Lambert, Jesús Giménez, Marta R. Costa-jussà, Enrique Amigó, Rafael E. Banchs, Lluís Màrquez and J.A. R. Fonollosa. **Machine Translation System Development Based on Human Likeness**. IEEE/ACL Workshop on Spoken Language Technology 2006

- 1 Introduction
- 2 N-gram-based Machine Translation
- 3 Word Alignment (Summary)
- 4 Multi-word Expression Grouping (Summary)
- 5 Parameter Optimisation Improvements (Summary)
- 6 Alignment Minimum-Translation-Error Training**
 - Introduction
 - Alignment System
 - Alignment Model Weight Optimisation
 - Experimental Work
- 7 Conclusions and Further Work

Motivation

⇒ alignment adapted to the SMT system

⇒ align quickly a short increment of parallel corpus

Related Work

- generative models:
 - not easily adaptable
 - large computational resources
- discriminative approach (log-linear combination of models):
 - easily adaptable to MT system (**not done in practise**)
 - some of them [Moore 2005] don't require heavy training
 - **inadequate model weight optimisation**
- problem with model weight optimisation:
 - need of data annotated with word alignment can be a limitation
 - main difficulty: absence of adequate metric (bad correlation AER-MT)
- new metrics proposed [Fraser and Marcu 2006; Ayan and Dorr 2006]: more informative for translation units extraction, but **indirect** method

Proposed Method

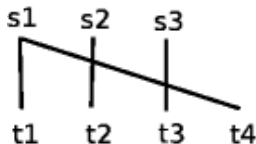
- Discriminative word alignment framework with some models **designed based on the characteristics of our SMT system**
- Optimise alignment model weights **without annotated data, directly** in a Minimum-translation-error Training scheme: use automated translation metrics as minimisation criterion.
- For large corpora, weights are optimised on a small part of the corpus and used to align whole corpus

discriminative framework, “structured” models (i.e. not local):

- easy to incorporate new features
- no observation sequence (*i.e.* manually annotated data) needed, just a score (*e.g.* MT metric)
- not globally optimal solution to the alignment problem (beam search). however: better alignments do not imply better translations .

BIA: Bilingual word Aligner

- Beam-search decoder minimising cost of linear combination of models (some of them similar to [Moore 2005])
- Two issues of N-gram MT can be dealt with at the alignment stage:
 - tuples with NULL source sides cannot be allowed
 - only one monotonic segmentation of each sentence pair is performed
⇒ long reorderings produce long and sparse tuples
- Features designed to help the N-gram SMT system:
 - distinct source and target unlinked word penalties
 - embedded word position penalty: penalises situations like this one:



- link bonus: because N-gram model prefers higher recall alignments

BIA (ii)

Other basic features:

- two distortion features: number and amplitude of crossing links
- word association model(s)

Second pass features:

- Word association model with relative link probabilities [Melamed 00]
- Fertility model: probability to have zero, one, two, three or four and more links

Association and Unlinked Word Models

- when possible, align stems instead of full forms
- word association model: χ^2 score like Cherry and Lin [03]?
Log-likelihood Ratio (LLR) like Moore [05] and Melamed [00]
(Dunning [93] showed that LLR accounts better for rare events) ?
- we also tried IBM1 probabilities: same order training time
- HMM or IBM4 probabilities: too costly for purpose of aligning rapidly a small corpus increment
- uniform unlinked word penalty can be replaced by a penalty proportional to the IBM1 NULL link probability

Association and Unlinked Word Models: results

Training data: proceedings of the European Parliament debates,
English-Spanish

Alignment development and test data: 245 sentences each

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1.28 M	34.9 M	106 k	27.2
Spa	1.28 M	36.6 M	153 k	28.5

Association and Unlinked Word Models: results

Training data: proceedings of the European Parliament debates,
English-Spanish

Alignment development and test data: 245 sentences each

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1.28 M	34.9 M	106 k	27.2
Spa	1.28 M	36.6 M	153 k	28.5

Average and standard error (in parentheses) of 3 alignment weight optimisations shown

Results on EPPS corpus	Rs	Pp	AER
χ^2 words	59.9 (0.3)	85.0 (0.4)	29.4 (0.1)
χ^2 stems	62.4 (0.8)	86.7 (1.5)	27.1 (0.1)

Association and Unlinked Word Models: results

Training data: proceedings of the European Parliament debates,
English-Spanish

Alignment development and test data: 245 sentences each

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1.28 M	34.9 M	106 k	27.2
Spa	1.28 M	36.6 M	153 k	28.5

Average and standard error (in parentheses) of 3 alignment weight optimisations shown

Results on EPPS corpus	Rs	Pp	AER
χ^2 words	59.9 (0.3)	85.0 (0.4)	29.4 (0.1)
χ^2 stems	62.4 (0.8)	86.7 (1.5)	27.1 (0.1)
LLR stems	59.4 (0.1)	75.7 (0.5)	33.2 (0.3)

Association and Unlinked Word Models: results

Training data: proceedings of the European Parliament debates,
English-Spanish

Alignment development and test data: 245 sentences each

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1.28 M	34.9 M	106 k	27.2
Spa	1.28 M	36.6 M	153 k	28.5

Average and standard error (in parentheses) of 3 alignment weight optimisations shown

Results on EPPS corpus	Rs	Pp	AER
χ^2 words	59.9 (0.3)	85.0 (0.4)	29.4 (0.1)
χ^2 stems	62.4 (0.8)	86.7 (1.5)	27.1 (0.1)
LLR stems	59.4 (0.1)	75.7 (0.5)	33.2 (0.3)
IBM1 stems	65.9 (0.7)	90.3 (1.4)	23.5 (0.3)

Association and Unlinked Word Models: results

Training data: proceedings of the European Parliament debates,
English-Spanish

Alignment development and test data: 245 sentences each

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1.28 M	34.9 M	106 k	27.2
Spa	1.28 M	36.6 M	153 k	28.5

Average and standard error (in parentheses) of 3 alignment weight optimisations shown

Results on EPPS corpus	Rs	Pp	AER
χ^2 words	59.9 (0.3)	85.0 (0.4)	29.4 (0.1)
χ^2 stems	62.4 (0.8)	86.7 (1.5)	27.1 (0.1)
LLR stems	59.4 (0.1)	75.7 (0.5)	33.2 (0.3)
IBM1 stems	65.9 (0.7)	90.3 (1.4)	23.5 (0.3)
IBM1+UM stems	67.1 (0.3)	92.5 (0.4)	21.9 (0.3)

Search (i): Possible links list

- List of links considered during search
- Obtained by pruning the association model table:
 - select best N target words for each source word
 - select best N source words for each target word

Sentence pair:

(0)the (1)member (2)state (3).

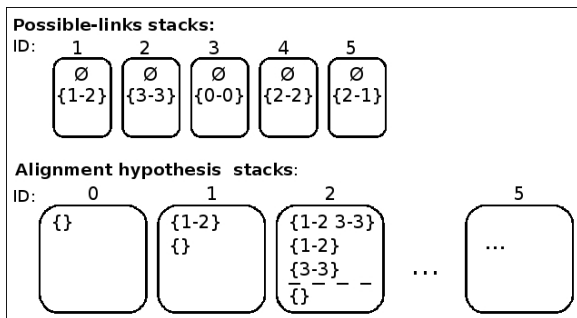
(0)los (1)pais (2)miembr (3).

Possible links ($N = 1$, association model cost order):

Link	Cost	Corresponding words
1-2	0.1736	member-miembr
3-3	0.6758	.-.
0-0	1.3865	the-los
2-2	1.8285	state-miembr
2-1	2.4027	state-pais

Search (ii): Baseline Search

- Possible links are arranged in link stacks to be expanded during search
- Moore's search [05]: baseline search



Search (iii): Problem with Baseline Search

- Main drawback: final alignment depends on the order in which links are introduced

the member state .
| \ |
los pays miembr .

Next possible links:

1.8 state-miembr

2.4 state-pays

Search (iii): Problem with Baseline Search

- Main drawback: final alignment depends on the order in which links are introduced

the	member	state	.
los	pais	miembr	.

Diagram illustrating a partial alignment between the English sentence "the member state ." and the Spanish sentence "los pais miembr .". Vertical lines connect "the" to "los" and "state" to "miembr". A diagonal line connects "member" to "pais".

Next possible links:

1.8 state-miembr
2.4 state-pais

Model weights:

1 word association (wa)
0.5 distortion (d)
2 unlinked word (um)

Search (iii): Problem with Baseline Search

- Main drawback: final alignment depends on the order in which links are introduced

<p>the member state .</p> <p> / </p> <p>los pais miembr .</p>	<table style="border-collapse: collapse;"> <tr><td>wa</td><td>1.8</td></tr> <tr><td>d</td><td>0</td></tr> <tr><td>um</td><td>-2</td></tr> <tr><td colspan="2"><hr style="border: 0; border-top: 1px solid black;"/></td></tr> <tr><td></td><td>-0.2</td></tr> </table>	wa	1.8	d	0	um	-2	<hr style="border: 0; border-top: 1px solid black;"/>			-0.2
wa	1.8										
d	0										
um	-2										
<hr style="border: 0; border-top: 1px solid black;"/>											
	-0.2										
<p><u>Next possible links:</u></p> <p>1.8 state-miembr</p> <p>2.4 state-pais</p>	<p><u>Model weights:</u></p> <p>1 word association (wa)</p> <p>0.5 distortion (d)</p> <p>2 unlinked word (um)</p>										

Search (iii): Problem with Baseline Search

- Main drawback: final alignment depends on the order in which links are introduced

<p>the member state .</p> <p> / </p> <p>los pais miembr .</p> <p style="text-align: center;"> / \ </p>	<table style="border-collapse: collapse; margin: 0 auto;"> <tr><td style="padding: 2px 5px;">wa</td><td style="padding: 2px 5px;">2.4</td></tr> <tr><td style="padding: 2px 5px;">d</td><td style="padding: 2px 5px;">0.5</td></tr> <tr><td style="padding: 2px 5px;">um</td><td style="padding: 2px 5px;">-2</td></tr> <tr><td colspan="2" style="border-top: 1px solid black; padding: 2px 5px;">+0.9</td></tr> </table>	wa	2.4	d	0.5	um	-2	+0.9	
wa	2.4								
d	0.5								
um	-2								
+0.9									
<p><u>Next possible links:</u></p> <p>1.8 state-miembr</p> <p>2.4 state-pais</p>	<p><u>Model weights:</u></p> <p>1 word association (wa)</p> <p>0.5 distortion (d)</p> <p>2 unlinked word (um)</p>								

Search (iii): Problem with Baseline Search

- Main drawback: final alignment depends on the order in which links are introduced

the	member	state	.
los	pais	miembr	.

Diagram illustrating a partial alignment between the English sentence "the member state ." and the Spanish sentence "los pais miembr .". Vertical lines connect "the" to "los", "member" to "miembr", and "state" to "pais". A diagonal line connects "state" to "miembr".

Next possible links:

1.8 state-miembr
2.4 state-pais

Model weights:

1 word association (wa)
0.5 distortion (d)
2 unlinked word (um)

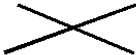
Search (iii): Problem with Baseline Search

- Main drawback: final alignment depends on the order in which links are introduced

<p>the member state .</p> <p> </p> <p>los pais miembr .</p>		<table border="0"> <tr><td>wa</td><td>2.4</td></tr> <tr><td>d</td><td>0.5</td></tr> <tr><td>um</td><td>-4</td></tr> <tr><td colspan="2"><hr/></td></tr> <tr><td></td><td>-1.1</td></tr> </table>	wa	2.4	d	0.5	um	-4	<hr/>			-1.1
wa	2.4											
d	0.5											
um	-4											
<hr/>												
	-1.1											
<p><u>Next possible links:</u></p> <p>1.8 state-miembr</p> <p>2.4 state-pais</p>	<p><u>Model weights:</u></p> <p>1 word association (wa)</p> <p>0.5 distortion (d)</p> <p>2 unlinked word (um)</p>											

Search (iii): Problem with Baseline Search

- Main drawback: final alignment depends on the order in which links are introduced

the	member	state	.
			
los	pais	miembr	.

Next possible links:

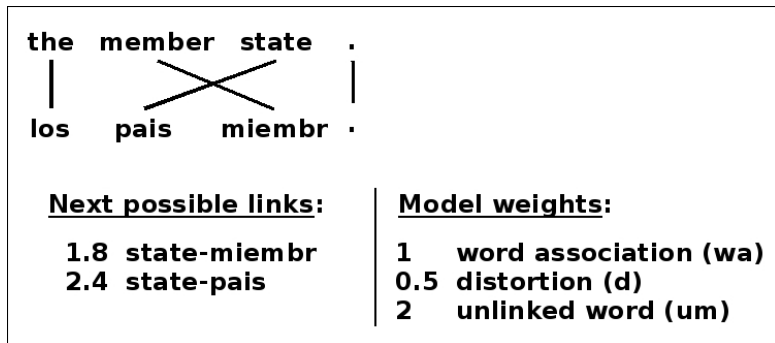
1.8 state-miembr
2.4 state-pais

Model weights:

1 word association (wa)
0.5 distortion (d)
2 unlinked word (um)

Search (iii): Problem with Baseline Search

- Main drawback: final alignment depends on the order in which links are introduced



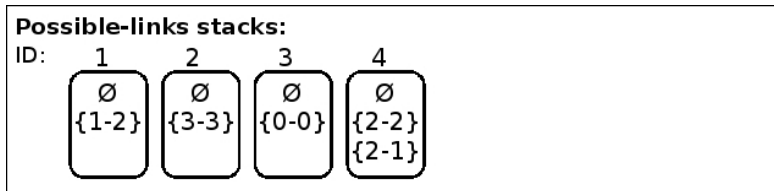
- Thus, probable but incorrect link introduced first prevented the correct link from being in final alignment
- Here, because of unlinked model; other cases: because of distortion

Search (iv): Improved Search

- Perform two successive iterations of alignment algorithm
 - second iteration: start from best alignment of first iteration
 - search more informed with context of first iteration links

Search (iv): Improved Search

- Perform two successive iterations of alignment algorithm
 - second iteration: start from best alignment of first iteration
 - search more informed with context of first iteration links
- Let links to same word compete on a fair basis



“source-word-score” (SWS) search strategy

- to allow many-to-many links: insert link sequences in stacks
- “source-word-position” (SWP) search strategy: follow source word position order

Search (iv): Improved Search

the member state .
| \ /
los pais miembr .

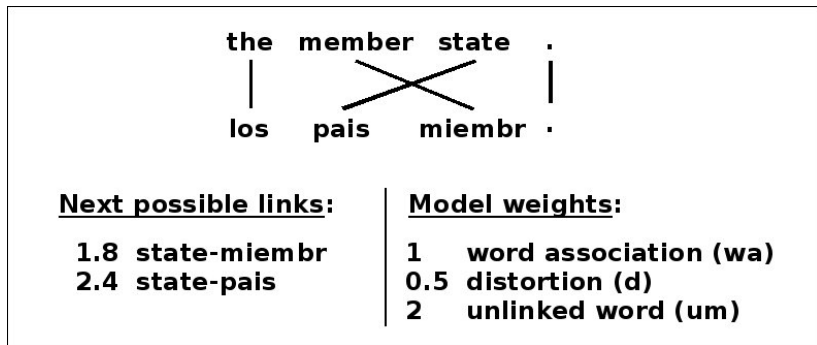
Next possible links:

1.8 state-miembr
2.4 state-pais

Model weights:

1 word association (wa)
0.5 distortion (d)
2 unlinked word (um)

Search (iv): Improved Search



- strategies following sentence words offer more flexibility to control complexity:
 - link stacks can be sorted and pruned out by histogram and/or threshold
 - number of links in the combinations can be restricted
 - restrict combinations to consecutive words

Search (v): Search Results

	Rs	Pp	AER
Baseline	67.1 (0.6)	93.3 (0.2)	21.6 (0.4)
SWP	66.9 (0.4)	92.0 (0.4)	22.2 (0.2)
SWS	67.8 (0.2)	93.7 (0.2)	21.0 (0.1)

- best strategy: following source words sorted according to their score

Final Alignment System

- source-word-score search, 3 possible associations per word ($N = 3$)
- First pass: $\hat{\mathbf{a}}^{(1)} = \arg \max_{\mathbf{a}} \sum \lambda_i^{(1)} h_i$

Feature functions:

- word association models based on source-target and target-source IBM1 probabilities
- link bonus; source and target unlinked word models proportional to IBM1 NULL link probabilities
- two distortion models (number and amplitude of crossing links)
- embedded word position penalty

Final Alignment System

- source-word-score search, 3 possible associations per word ($N = 3$)

- First pass: $\hat{\mathbf{a}}^{(1)} = \arg \max_{\mathbf{a}} \sum \lambda_i^{(1)} h_i$

Feature functions:

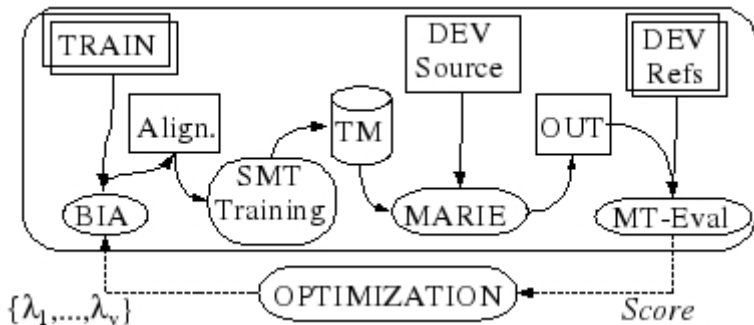
- word association models based on source-target and target-source IBM1 probabilities
- link bonus; source and target unlinked word models proportional to IBM1 NULL link probabilities
- two distortion models (number and amplitude of crossing links)
- embedded word position penalty
- Second pass: $\hat{\mathbf{a}}^{(2)} = \arg \max_{\mathbf{a}} \sum \lambda_i^{(2)} h_i$
 - word association model **with relative link probabilities**
 - link bonus; **source and target fertility models**
 - two distortion models (number and amplitude of crossing links)
 - embedded word position penalty

Procedure

- A small subset of the parallel corpus was selected, based on 2 keys:
 - ① number of occurrences in the training corpus of the least frequent word in the sentence pair
 - ② same criterion only among words present in the development data set

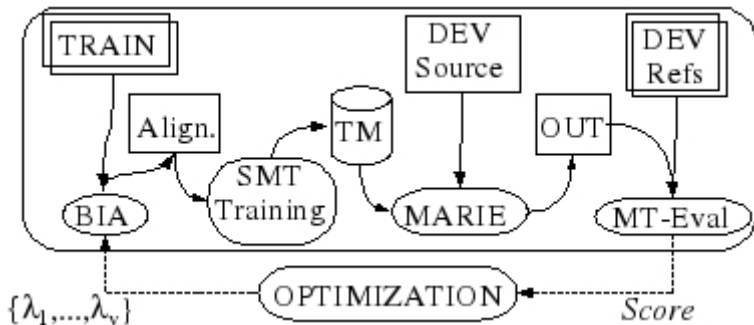
Procedure

- A small subset of the parallel corpus was selected, based on 2 keys:
 - ① number of occurrences in the training corpus of the least frequent word in the sentence pair
 - ② same criterion only among words present in the development data set
- Optimal coefficients were estimated **on small subset** as follows:



Procedure

- A small subset of the parallel corpus was selected, based on 2 keys:
 - ① number of occurrences in the training corpus of the least frequent word in the sentence pair
 - ② same criterion only among words present in the development data set
- Optimal coefficients were estimated **on small subset** as follows:



- variant: full N-gram SMT system used to translate DEV source

Optimisation Algorithm

- SPSA algorithm used.

Objective function: $BLEU(\lambda_1^{(1)}, \dots, \lambda_7^{(1)}, \lambda_1^{(2)}, \dots, \lambda_6^{(2)})$

- Each time parameters vary, this function is evaluated, which implies:
 - aligning selected sub-corpus with new set of parameters
 - extracting translation units
 - training translation model
 - translating dev corpus
 - evaluating translation \rightarrow score=value of objective function

Basic Experiment

- Once a set of optimal weights has been obtained:
 - align **whole** training corpus with optimal weights
 - extract translation units
 - train **full** SMT system: translation model+target language model, word bonus model and two lexical models.

Basic Experiment

- Once a set of optimal weights has been obtained:
 - align **whole** training corpus with optimal weights
 - extract translation units
 - train **full** SMT system: translation model+target language model, word bonus model and two lexical models.
- This SMT system was compared to identical system trained from combinations of source-target and target-source GIZA++ (50 cl, 1-4 H-5 4-4) alignments

Data Sets

- Ar→En United Nations

- training data:

Lang.	Sentences	Words	Vocab.	Aver.
Ar	1.43 M	45.2 M	191 k	31.7
Eng	1.43 M	43.8 M	134 k	30.7

- development (2 sets) and test: from NIST 2002, 2004 and 2005 evaluations

Data Sets

- Ar→En United Nations

- training data:

Lang.	Sentences	Words	Vocab.	Aver.
Ar	1.43 M	45.2 M	191 k	31.7
Eng	1.43 M	43.8 M	134 k	30.7

- development (2 sets) and test: from NIST 2002, 2004 and 2005 evaluations

- En→Es EPPS

- training data

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1.28 M	34.9 M	106 k	27.2
Spa	1.28 M	36.6 M	153 k	28.5

- development (2 sets) and test: 26k, 18.7k and 26.9k words, 2 references

AR→EN United Nations Results on 50k subset

- alignment weights tuned on 25k and 50k subset (25k included in 50k)
- Results shown are the average and standard error (in parentheses) of 3 **SMT** model weight optimisations

Results on 50k subset (**only difference between systems lies in alignment**):

	BLEU	NIST	WER	METEOR
25 BM	23.2 (0.1)	7.76 (0.03)	66.6 (0.4)	52.8 (0.1)
25 full	23.2 (0.5)	7.59 (0.2)	68.5 (2.5)	52.8 (0.5)
50 BM	23.8 (0.1)	7.77 (0.1)	67.1 (1.3)	53.2 (0.2)
50 full	23.5 (0.2)	7.69 (0.09)	67.6 (1.1)	52.9 (0.2)
Giza++ U	21.4 (0.4)	7.15 (0.13)	71.8 (2.2)	51.4 (0.4)
Giza++ GDF	22.5 (0.7)	7.41 (0.2)	70.7 (2.4)	52.9 (0.4)

AR→EN United Nations Results on full corpus

Results on full corpus (alignment weights tuned on 25k and 50k subset):

	BLEU	NIST	WER	METEOR
25 BM	26.0 (0.2)	8.14 (0.1)	64.3 (1.2)	54.0 (0.5)
25 full	26.3 (0.1)	8.18 (0.15)	63.6 (1.4)	53.6 (0.4)
50 BM	27.0 (0.2)	8.15 (0.07)	64.8 (0.8)	54.4 (0.2)
50 full	26.7 (0.3)	8.19 (0.06)	63.8 (0.7)	53.9 (0.4)
Giza++ U	26.0 (0.2)	8.10 (0.2)	64.8 (2.4)	53.5 (0.7)
Giza++ GDF	26.9 (0.3)	8.09 (0.1)	64.9 (1.0)	54.4 (0.1)

- Erratum: Giza++ training time for this corpus was 28 hours

EPPS Results

English→Spanish automatic translation results on full EPPS Corpus:

	BLEU	NIST	WER
25 BM	48.8 (0.3)	9.81 (0.08)	40.6 (0.6)
25 full	49.3 (0.2)	9.88 (0.01)	40.3 (0.06)
Giza++ U	49.1 (0.1)	9.83 (0.02)	40.9 (0.2)
Giza++ GDF	49.5 (0.3)	9.84 (0.04)	40.6 (0.3)

Selected Publications: Discriminative Alignment Training

- Patrik Lambert, Rafael Banchs and Josep M. Crego. **Discriminative Alignment Training without Annotated Data for Machine Translation.** NAACL-HLT (short papers) 2007

- 1 Introduction
- 2 N-gram-based Machine Translation
- 3 Word Alignment (Summary)
- 4 Multi-word Expression Grouping (Summary)
- 5 Parameter Optimisation Improvements (Summary)
- 6 Alignment Minimum-Translation-Error Training
- 7 Conclusions and Further Work**

Conclusions

- Guidelines for alignment evaluation and manual alignment
 - impact of evaluation procedure on scores, especially influence of S/P ratio on AER
- Experiments on linguistic classification to improve word alignment:
⇒ no correlation between AER and automatic MT metrics
- SMT quality can be improved:
 - by grouping MWEs before alignment
 - by tuning system according to more reliable criterion than BLEU (IQmt)
 - by tuning system with SPSA algorithm instead of Downhill Simplex

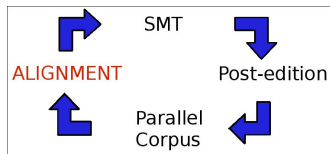
Conclusions

Discriminative Alignment Training:

- Implemented Discriminative Word Alignment System (BIA):
 - some models designed according to our SMT system characteristics
 - association model: IBM1 probabilities better than association measures
 - IBM1 NULL probabilities are better than uniform unlinked penalty
 - improved search
- Framework proposed: model weights tuned directly according to MT metrics, without alignment reference
 - small corpora: tuning on whole corpus. Better results than Giza++
 - large corpora: tuning on small corpus subset. Weights useful to align whole corpus (results as good as Giza++)
- Limitations: sensitive to optimisation process; weights tuning time

Conclusions

- Flexible alignment technique:
 - very low average memory usage is required
 - straightforward to run in parallel
 - quality alignments obtained very quickly for small corpus updates (if weights known)
- ⇒ this loop can be performed more often:



Further Work

Multi-word Expressions:

- refine extraction method
- use manually built resources (wordnet?)

Parameter Optimisation:

- use more systematically QUEEN metric of IQmt framework
- evaluate performance of SPSA algorithm in higher dimensionality

Discriminative Alignment Training:

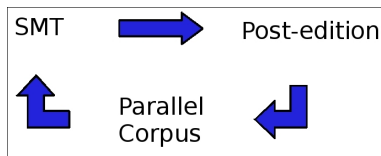
- improve the alignment system
- analyse the differences between alignments produced after optimisation in function of AER or MT metrics
- study the effect of the size of the subcorpus used to train alignment parameters, as well as the effect of the sentences selection method.

Thank you for your attention !

Thank you for your attention !

The Statistical Approach to Machine Translation (ii)

- Advantages
 - reduced effort (system built automatically from parallel corpus)
 - reduced required skills (no knowledge of language required)
- Drawbacks for commercial systems:
 - very domain-dependent
 - not easy to correct errors (require to modify models or corpus)
 - not easy to customise
- Framework to alleviate these drawbacks:



- adapt to user domain
- correct some errors
- customise corpus to user needs

From the translation model to the translation system

- n-gram-based translation model alone can produce good translations, but search is better guided with more models
- $\hat{T} = \arg \max_T \sum_m \lambda_m h_m(S, T)$; Features:
 - Target language model (standard n-gram model)
 - Word bonus model: $p_{WP}(T) = \exp(\text{number of words in } T)$.

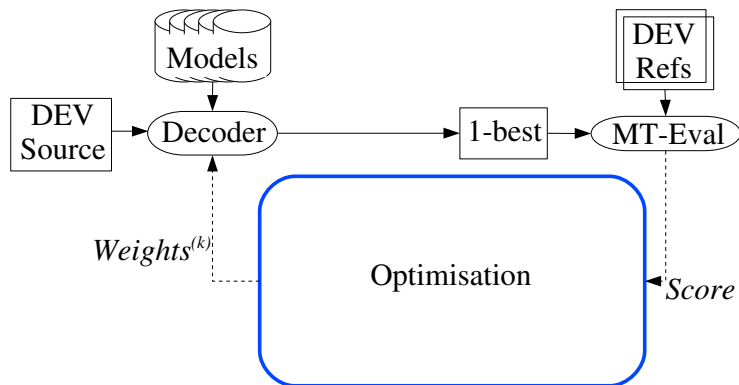
From the translation model to the translation system

- n-gram-based translation model alone can produce good translations, but search is better guided with more models
- $\hat{T} = \arg \max_T \sum_m \lambda_m h_m(S, T)$; Features:
 - Target language model (standard n-gram model)
 - Word bonus model: $p_{WP}(T) = \exp(\text{number of words in } T)$.
 - A source-target lexical model, which use IBM1 translation probabilities to compute a lexical weight for each tuple

$$p_{IBM1}((\tilde{s}, \tilde{t})_k) = \frac{1}{(I+1)^J} \prod_{j=1}^J \sum_{i=0}^I p(t_k^i | s_k^j)$$

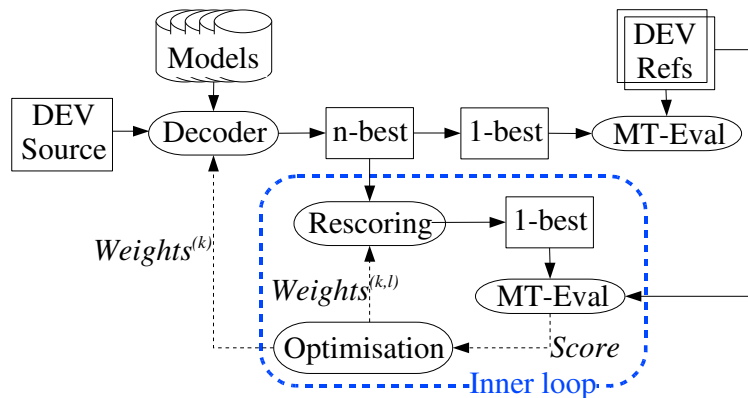
- A target-source lexical model
- Model weight optimisation: implementation of a tool based on two possible algorithms: Downhill Simplex and SPSA

Model Weight Optimisation Procedures



- Single loop: \approx 50 to 100 decodings and MT evaluations

Model Weight Optimisation Procedures



- Single loop: \approx 50 to 100 decodings and MT evaluations
- **outer loop**: 5-10 decodings and evaluations;
- **inner loop**: \approx 50 re-rankings (and evaluations).

AR→EN United Nations Results on full corpus

Results on full corpus (alignment weights tuned on 25k and 50k subset):

	BLEU	NIST	WER	METEOR
1 pass	24.5 (0.3)	7.87 (0.06)	66.1 (0.8)	53.1 (0.4)
1 iter	24.1 (0.1)	7.76 (0.06)	67.8 (0.7)	53.8 (0.2)
25 BM	26.0 (0.2)	8.14 (0.1)	64.3 (1.2)	54.0 (0.5)
25 full	26.3 (0.1)	8.18 (0.15)	63.6 (1.4)	53.6 (0.4)
50 BM	27.0 (0.2)	8.15 (0.07)	64.8 (0.8)	54.4 (0.2)
50 full	26.7 (0.3)	8.19 (0.06)	63.8 (0.7)	53.9 (0.4)
Giza++ U	26.0 (0.2)	8.10 (0.2)	64.8 (2.4)	53.5 (0.7)
Giza++ GDF	26.9 (0.3)	8.09 (0.1)	64.9 (1.0)	54.4 (0.1)

EPPS Results

English→Spanish automatic translation results on full EPPS Corpus:

	AER	BLEU	NIST	WER
25 BM	20.9 / 18.9	48.8 (0.3)	9.81 (0.08)	40.6 (0.6)
25 full	23.0 / 19.1	49.3 (0.2)	9.88 (0.01)	40.3 (0.06)
Giza++ U	15.2	49.1 (0.1)	9.83 (0.02)	40.9 (0.2)
Giza++ GDF	14.4	49.5 (0.3)	9.84 (0.04)	40.6 (0.3)

Further Work

Discriminative Alignment Training:

- improve the alignment system
 - systematic study of impact of each parameter and feature in translation
 - in the second pass, calculate probabilities for links between phrases
 - include MWE detection
 - try estimation of remaining cost in search
 - add features (for example syntax-based)
- improve model weight optimisation procedure (control of over-tuning on a second set)
- analyse the differences between alignments produced after optimisation in function of AER or MT metrics
- study the effect of the size of the subcorpus used to train alignment parameters, as well as the effect of the sentences selection method.

Thank you for your attention !