

Modeling Relevance in Statistical MT

*Scoring Alignment, Context, and Annotations of
Translation Instances*

Aaron B. Phillips

Language Technologies Institute
Carnegie Mellon University

January 26th, 2012
Thesis Defense

Outline

1 Background & Motivation

2 Cunei Machine Translation Platform

- Baseline: Modeling Phrase Alignment
- Extension 1: Modeling Source Similarity
- Extension 2: Modeling Target Similarity
- Extension 3: Incorporating Corpus Annotations

3 Conclusions

Outline

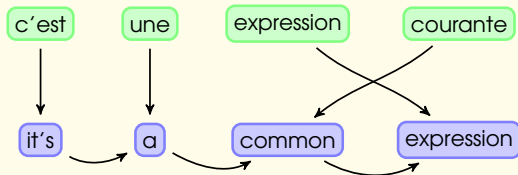
1 Background & Motivation

2 Cunei Machine Translation Platform

- Baseline: Modeling Phrase Alignment
- Extension 1: Modeling Source Similarity
- Extension 2: Modeling Target Similarity
- Extension 3: Incorporating Corpus Annotations

3 Conclusions

Statistical Modeling in MT



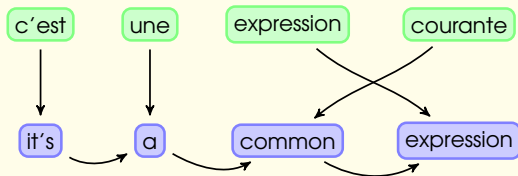
Step 1 Select what **units** to model

Step 2 Select how to **score** each translation unit

Step 3 Select how to **combine** translation units

Standard Modeling Approach

Translation Model $P(s|t)$ $P(t|s)$ $lex(s|t)$ $lex(t|s)$



Language Model $P(t_3|t_1 t_2)$

- Log-linear model with multiple features
- Typically features are relative frequency estimates
- Model new information with conditional likelihoods

Domain Sensitivity

In-domain text is text that is generated by the model on the same domain as the training data. This text is used to evaluate the model's performance on the domain it was trained on.

Out-of-domain text is text that is generated by the model on a different domain than the training data. This text is used to evaluate the model's performance on new domains.

The model's performance is generally higher on in-domain text than on out-of-domain text. This is because the model has seen the in-domain text during training and has learned to generate text that is similar to it.

To evaluate the model's performance on out-of-domain text, we need to use a separate set of data. This data should be generated by the model on a domain that is different from the training domain.

In-Domain Text

Out-of-domain text is text that is generated by the model on a different domain than the training data. This text is used to evaluate the model's performance on new domains.

The model's performance is generally lower on out-of-domain text than on in-domain text. This is because the model has not seen the out-of-domain text during training and has not learned to generate text that is similar to it.

To evaluate the model's performance on out-of-domain text, we need to use a separate set of data. This data should be generated by the model on a domain that is different from the training domain.

Out-of-Domain Text

Compute likelihood conditioned on being in-domain

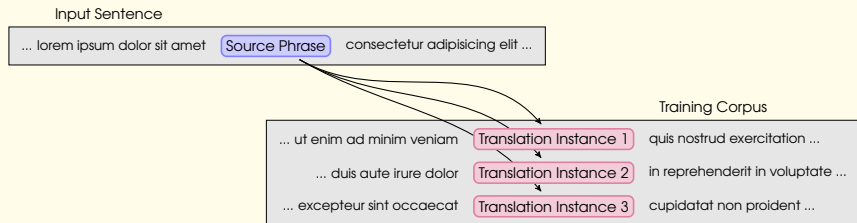
- Trade-off between bias and variance
- Learn appropriate weights during training

$$\begin{array}{cccc}
 P(s|t) & P(t|s) & lex(s|t) & lex(t|s) \\
 \hline
 P(s|t, d) & P(t|s, d) & lex(s|t, d) & lex(t|s, d)
 \end{array}$$

The Problem

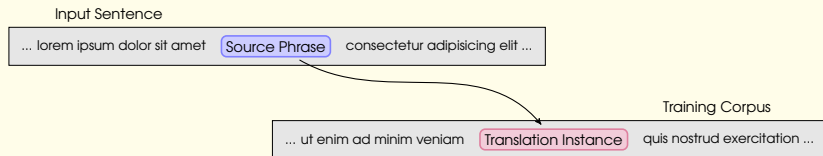
- We cannot model all possible dependencies (the number of features quickly becomes untenable)
 - Often features selection is based on heuristics, intuition, and trial-and-error
- It is difficult to inject the notion of **relevance**
 - Relative frequency estimates typically assume that all evidence is equal
 - We can marginalize over additional information, but the distribution(s) must be decided on a priori

Modeling Translation Instances



Instance of Translation - the realization of a source and target pair at *one specific location in the corpus*

Modeling Translation Instances



Information Associated with each Instance of Translation

- Document Context (Genre)
- Local Sentential Context
- Phrase Alignment
- Consistency of Annotations
- Target-Side Context

Thesis Statement

Modeling each instance of a translation in the corpus will improve machine translation quality and facilitate the integration of non-local context and similarity features

Outline

1 Background & Motivation

2 Cunei Machine Translation Platform

- Baseline: Modeling Phrase Alignment
- Extension 1: Modeling Source Similarity
- Extension 2: Modeling Target Similarity
- Extension 3: Incorporating Corpus Annotations

3 Conclusions

Formalism

- Standard Decision Rule used in Machine Translation

$$\tilde{t} = \arg \max_{t_1, t_2 \dots t_n} \sum_{i=0}^n m(s_i, t_i, \lambda)$$

- Model used in Statistical Machine Translation

$$\begin{aligned} m(s_i, t_i, \lambda) &= \sum_k \lambda_k \cdot \theta_k(s_i, t_i) \\ &= \ln e^{\sum_k \lambda_k \cdot \theta_k(s_i, t_i)} \end{aligned}$$

- Model used by Cunei

$$m(s_i, t_i, \lambda) = \ln \sum_{\eta} e^{\sum_k \lambda_k \cdot \phi_k(s_i, t_i, \eta)}$$

Formalism

- Standard Decision Rule used in Machine Translation

$$\tilde{t} = \arg \max_{t_1, t_2 \dots t_n} \sum_{i=0}^n m(s_i, t_i, \lambda)$$

- Model used in Statistical Machine Translation

$$\begin{aligned} m(s_i, t_i, \lambda) &= \sum_k \lambda_k \cdot \theta_k(s_i, t_i) \\ &= \ln e^{\sum_k \lambda_k \cdot \theta_k(s_i, t_i)} \end{aligned}$$

- Model used by Cunei

$$m(s_i, t_i, \lambda) = \ln \sum_{\eta} e^{\sum_k \lambda_k \cdot \phi_k(s_i, t_i, \eta)}$$

Formalism

- Standard Decision Rule used in Machine Translation

$$\tilde{t} = \arg \max_{t_1, t_2 \dots t_n} \sum_{i=0}^n m(s_i, t_i, \lambda)$$

- Model used in Statistical Machine Translation

$$\begin{aligned} m(s_i, t_i, \lambda) &= \sum_k \lambda_k \cdot \theta_k(s_i, t_i) \\ &= \ln e^{\sum_k \lambda_k \cdot \theta_k(s_i, t_i)} \end{aligned}$$

- Model used by Cunei

$$m(s_i, t_i, \lambda) = \ln \sum_{\eta} e^{\sum_k \lambda_k \cdot \phi_k(s_i, t_i, \eta)}$$

Formalism

- Standard Decision Rule used in Machine Translation

$$\tilde{t} = \arg \max_{t_1, t_2 \dots t_n} \sum_{i=0}^n m(s_i, t_i, \lambda)$$

- Model used in Statistical Machine Translation

$$\begin{aligned} m(s_i, t_i, \lambda) &= \sum_k \lambda_k \cdot \theta_k(s_i, t_i) \\ &= \ln e^{\sum_k \lambda_k \cdot \theta_k(s_i, t_i)} \end{aligned}$$

- Model used by Cunei

$$m(s_i, t_i, \lambda) = \ln \sum_{\eta} e^{\sum_k \lambda_k \cdot \phi_k(s_i, t_i, \eta)}$$

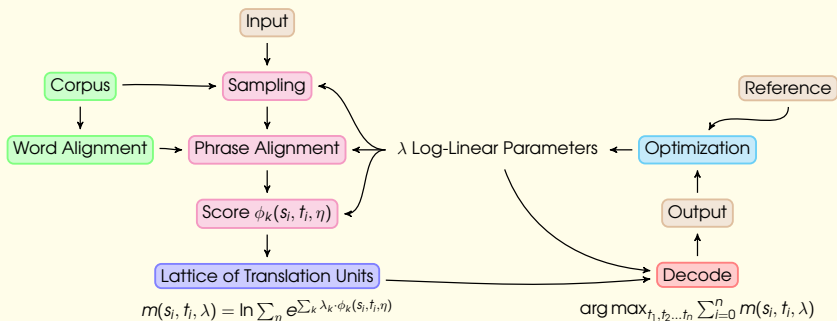
Relationship with SMT

If the features for all translation instances are constant

$$\phi_k(s, t, \eta) = \theta_k(s, t) \forall \eta, k$$

Then Cunei's model simplifies to the standard SMT model

System Architecture



Learning Model Weights

- Complicated by the fact that the score for each translation instance is dependent on λ
 - Use a second-order Taylor series to approximate the score of $m(s, t, \lambda)$ from $m(s, t, \lambda')$
 - Merge the n-best lists after each iteration
 - Discount models based on the distance from λ to λ'
- Built-in training follows (Smith and Eisner, 2006)'s annealing method to maximize $\log \mathbb{E}[\text{BLEU}]$

Advantages

- Easy to model features dependent on the particular translation instance, input, or surrounding translations
 - Knowledge is non-local to traditional SMT phrase pairs
- Efficiently search a very large hypothesis space
 - Postpone most modeling decisions until run-time
 - Use any information in the corpus for scoring the relevance of a translation instance
- The same model identifies and scores translations

Outline

1 Background & Motivation

2 Cunei Machine Translation Platform

- Baseline: Modeling Phrase Alignment
- Extension 1: Modeling Source Similarity
- Extension 2: Modeling Target Similarity
- Extension 3: Incorporating Corpus Annotations

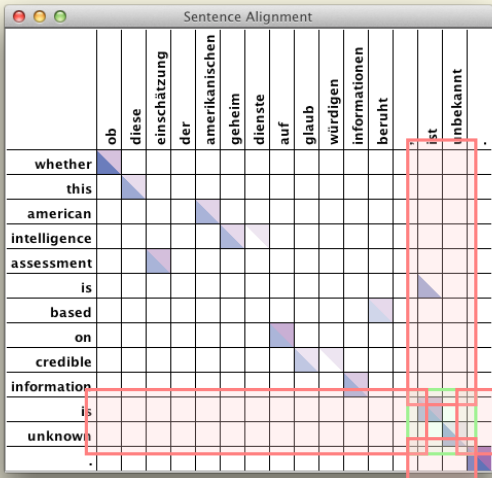
3 Conclusions

Phrase Alignment in Moses

- Uses a heuristic over the word alignments to determine a binary phrase alignment
- A phrase-pair will not be aligned if any word of the phrase-pair aligns elsewhere in the sentence

Phrase Alignment in Cunei

- Use word alignments as features for an on-line phrase alignment (Vogel, 2005)
- Not all instances of the translation will receive the same alignment score



Evaluation Method

German-English

- 100 million words from Europarl and WMT 2011 newswire
- Development and test sets from Europarl

Czech-English

- 40 million words (sampled uniformly) from CzEng 0.9 and WMT 2011 newswire
- Development and test sets from CzEng 0.9 (sampled by genre)

English language model trained on 512 million words

Moses vs Cunei

German-English

	BLEU	NIST	Meteor	TER
<i>Moses</i>	0.2534	6.6090	0.5185	0.5995
<i>Cunei</i>	0.2576 (1.66%)	6.6753 (1.00%)	0.5213 (0.54%)	0.5945 (0.83%)

Czech-English

	BLEU	NIST	Meteor	TER
<i>Moses</i>	0.2709	6.8378	0.4948	0.5704
<i>Cunei</i>	0.3076 (13.55%)	7.2122 (5.48%)	0.5249 (6.08%)	0.5385 (5.59%)

German Europarl Test Sentence #311

Moses that is exactly what has happened
in **the former yugoslav republic of**
macedonia .

Cunei that is exactly what happened in
macedonia .

Reference that is exactly what has happened in
macedonia .

Outline

1 Background & Motivation

2 Cunei Machine Translation Platform

- Baseline: Modeling Phrase Alignment
- Extension 1: Modeling Source Similarity
- Extension 2: Modeling Target Similarity
- Extension 3: Incorporating Corpus Annotations

3 Conclusions

The Role of Context

Definition

context *n.* the parts of a discourse that surround a word or passage and can throw light on its meaning
(Merriam-Webster)

Permits a more nuanced differentiation between each translation instance present in the corpus

Types of Context

Context from Sentence Annotations

- Static
- Dynamic

Context from Surrounding Tokens

- Sentence
- Document

Sentence Annotations

The Europarl distribution includes XML markup containing additional information about the text

One such sentence was...

- recorded in the **Europarl** proceedings
- in **November**
- of the year **2003**
- spoken originally in **Spanish**
- by **Vice-President of the Commission**
- with the name **De Palacio**

Example of Sentence Annotations

Input Sentence

i tipped the cab driver and he drove away

{ Genre : Fiction
Document : smith-173-08
Language : English
Year : 1999

Corpus Sentence for Translation Instance #1

she was talking to the cab driver .

{ Genre : Fiction
Document : brown-1274
Language : English
Year : 1999

Corpus Sentence for Translation Instance #2

if you have a disk that contains the updated driver , click ok .

{ Genre : Technical
Document : msdn-841
Language : English
Year : 2003

Context from Sentence Annotations

Dynamic Annotation Features

- One feature for each type of annotation (genre, author, year, etc.)
- Compute accuracy between the set of values associated with the annotation on the translation instance and the input

Static Annotation Features

- A mixture model over all annotation-defined collections that exist in the corpus
- Most appropriate when the development set closely matches the test set

Example of Surrounding Tokens

Input Sentences

after **retrieving** a newspaper i flagged down a ride across town
the **taxi** dropped me off at the **turnaround**
i tipped **the cab** **driver** and he drove away
it was then that i remembered my briefcase was still in the **car**

Translation Instance #1 with Corpus Context

the **taxi** pulled into the **turnaround** of the hotel .
he saw meredith 's **car** up ahead .
she was talking to **the cab** **driver** .
she looked back and saw him .

Translation Instance #2 with Corpus Context

retrieving a list of all devices
windows was unable to find any drivers for this device .
if you have a disk that contains the updated **driver** , click ok .
do you want to continue installing this driver ?

Context from Surrounding Tokens

Document Context Features

- Each document is modeled as a bag of words
- Compute cosine distance, Jensen-Shannon distance, precision, and recall as features
- Can be calculated over actual document boundaries or windows of sentences (or both)

Sentential Context Features

- Independently score left and right contexts
- Binary 1-gram, 2-gram, and 3-gram match features

Source Context with German Europarl v6

	BLEU	NIST	Meteor	TER
<i>Baseline</i>	0.2576	6.6753	0.5213	0.5945
<i>+ Static Annotations</i>	0.2650	6.7346	0.5222	0.5913
<i>+ Dynamic Annotations</i>	0.2617	6.6988	0.5217	0.5950
<i>+ Sentence Context</i>	0.2663	6.7636	0.5236	0.5882
<i>+ Document Context</i>	0.2622	6.7379	0.5230	0.5914
<i>All Context Features</i>	0.2686 (4.27%)	6.7668 (1.37%)	0.5214 (0.02%)	0.5862 (1.40%)

Source Context with CzEng v0.9

	BLEU	NIST	Meteor	TER
<i>Baseline</i>	0.3076	7.2122	0.5249	0.5385
<i>+ Static Annotations</i>	0.3077	7.2106	0.5244	0.5380
<i>+ Dynamic Annotations</i>	0.3101	7.2413	0.5254	0.5351
<i>+ Sentence Context</i>	0.3091	7.1994	0.5260	0.5381
<i>+ Document Context</i>	0.3105	7.2463	0.5291	0.5345
<i>All Context Features</i>	0.3120 (1.43%)	7.2708 (0.81%)	0.5290 (0.78%)	0.5321 (1.19%)

CzEng Test Sentence #449

<i>Baseline</i>	the % 1 service announced invalid the status quo % 2 .
<i>+ Static Annotations</i>	... announced invalid the current state % 2 .
<i>+ Dynamic Annotations</i>	... announced invalid the current state % 2 .
<i>+ Sentence Context</i>	... announced invalid the status quo % 2 .
<i>+ Document Context</i>	... announced invalid state of play % 2 .
<i>All Context Features</i>	... announced invalid the current state % 2 .
<i>Reference</i>	the % 1 service has reported an invalid current state % 2 .

Outline

1 Background & Motivation

2 Cunei Machine Translation Platform

- Baseline: Modeling Phrase Alignment
- Extension 1: Modeling Source Similarity
- Extension 2: Modeling Target Similarity
- Extension 3: Incorporating Corpus Annotations

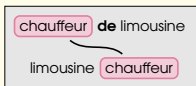
3 Conclusions

Context Available in Source and Target

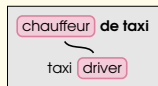
Input Sentence

où est le chauffeur de taxi ?

Corpus Sentence for Translation Instance #1



Corpus Sentence for Translation Instance #2

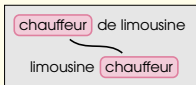


Context Available in Source and Target

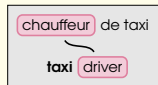
Input Sentence

où est le chauffeur de taxi ?

Corpus Sentence for Translation Instance #1



Corpus Sentence for Translation Instance #2



Output Sentence

where is the taxi

Limitations of Target Context

The output sentence is not completely known
(unlike the input sentence)

- Document context is too expensive
- Compare **left context** from the translation instance with the partially-constructed output
- Binary 1-gram, 2-gram, and 3-gram match features

(Annotations are the same for the source and target)

Target Context vs Language Modeling

Both aim to reduce boundary friction and improve fluency

The target context score ...

- is dependent on the source phrase
- uses translation instances weighted by source context, alignment probability, and all other features
- instead of smoothing, has features for each n -gram

Target Context

German-English

	BLEU	NIST	Meteor	TER
<i>Baseline</i>	0.2576	6.6753	0.5213	0.5945
<i>+Target Context</i>	0.2595 (0.74%)	6.6778 (0.04%)	0.5215 (0.04%)	0.5943 (0.03%)

Czech-English

	BLEU	NIST	Meteor	TER
<i>Baseline</i>	0.3076	7.2122	0.5249	0.5385
<i>+Target Context</i>	0.3102 (0.85%)	7.2282 (0.22%)	0.5244 (-0.10%)	0.5375 (0.19%)

CzEng Test Sentence #1348

Baseline

because the french use **the large**
roman numerals , when refer to the

+ Target Context

because the french use **capital**
roman numerals , when refer to the

Reference

since the french use capital roman
numerals to refer to the

Outline

1 Background & Motivation

2 Cunei Machine Translation Platform

- Baseline: Modeling Phrase Alignment
- Extension 1: Modeling Source Similarity
- Extension 2: Modeling Target Similarity
- Extension 3: Incorporating Corpus Annotations

3 Conclusions

The Role of Annotations

Definition

annotation *n.* a note added by way of comment or explanation (Merriam-Webster)

- May be created by humans or with ML algorithms
- May describe a document, sentence, or token
- May be present on the source-side and/or the target-side of the parallel corpus

Types of Annotations

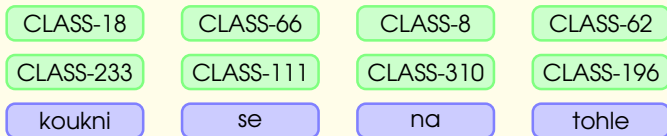
Sequential Annotation Labels

- Annotation that labels each word in the corpus
- Indexed as a type sequence which enables search

Hierarchical Annotations

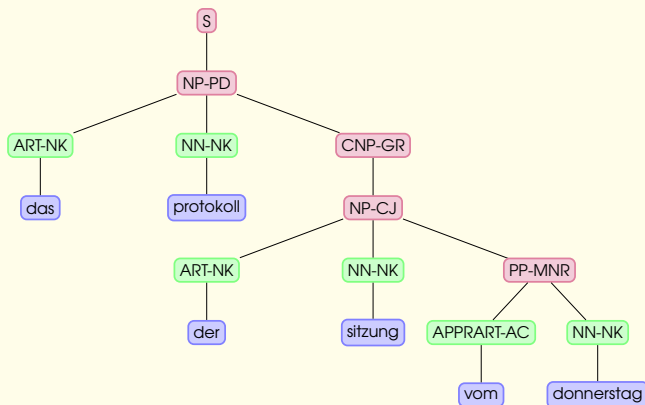
- Allows annotations to span multiple words
- Each annotation optionally references a parent

Czech-English Annotations



- Automatically create sequential annotation labels using MKCLS for unsupervised learning (Och, 1999)
- Two levels of granularity: 100 and 1000 clusters

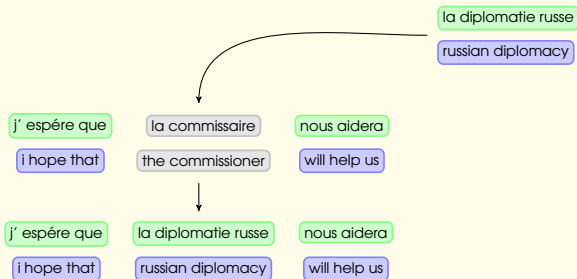
German-English Annotations



- Used the Stanford parser and built-in factored models to independently parse German and English

Replacement

Sequential annotations enable retrieval of translation instances that are *lexically divergent* from the input



Scoring Annotations

Purpose of annotations is to better model the relevance of each translation instance

Similarity Features

- Input Similarity (Source)
- Replacement Similarity (Target)

Extend Existing Features

- Source Context
- Translation Probability
- Target Context

Experiments

Annotations without Lexical Divergences

Same lexical hypotheses as the baseline system, but the translation model is augmented with annotation features

Annotations with Divergences

Allows translation instances that do not lexically match the input if they match one (or more) annotation sequences

Annotations with Divergences and Replacement

Allows part of a hypothesis to be replaced when it diverges from the input

Annotations with German Europarl v6

	BLEU	NIST	Meteor	TER
<i>Baseline</i>	25.76	6.675	52.13	59.45
<i>+Annotations without Lexical Divergences</i>	26.06	6.604	51.91	59.76
<i>+Annotations with Divergences</i>	26.08	6.644	52.06	59.60
<i>+Annotations with Divergences and Replacement</i>	26.15 (1.51%)	6.641 (-0.51%)	51.96 (-0.33%)	59.40 (0.08%)

Annotations with CzEng v0.9

	BLEU	NIST	Meteor	TER
<i>Baseline</i>	30.76	7.212	52.49	53.85
<i>+Annotations without Lexical Divergences</i>	32.85	7.362	53.29	52.59
<i>+Annotations with Divergences</i>	32.50	7.319	53.07	52.74
<i>+Annotations with Divergences and Replacement</i>	32.87 (6.86%)	7.354 (1.97%)	53.47 (1.87%)	52.68 (2.17%)

CzEng Test Sentence #719

<i>Baseline</i>	- article 4 of the agreement bulgaria - spain
<i>+ Annotations without Lexical Divergence</i>	- article 4 of the bulgaria - spain
<i>+ Annotations with Divergences</i>	- article 4 of the morocco - spain agreement ;
<i>+ Annotations with Divergences and Replacement</i>	- article 4 of the bulgaria - spain
<i>Reference</i>	- article 4 of the bulgaria - spain agreement ;

Outline

1 Background & Motivation

2 Cunei Machine Translation Platform

- Baseline: Modeling Phrase Alignment
- Extension 1: Modeling Source Similarity
- Extension 2: Modeling Target Similarity
- Extension 3: Incorporating Corpus Annotations

3 Conclusions

Contributions

Cunei's model allows **adaptation at the level of the translation unit** by scoring instances of translation

- Phrase Alignment
- Source Similarity
- Target Similarity
- Corpus Annotations

Related Work

- Build mixture of multiple translation models (Foster and Kuhn, 2007, Lu et al., 2007)
- Weight corpus documents based on similarity to the input (Hildebrand et al., 2005, Lu et al., 2007)
- Learn sentence weights based on a development set (Shah et al., 2010, Matsoukas et al., 2009)

Unique to Our Work

- Our features are more specific in that they operate over translation instances and not just sentences
- We construct **a single unified model** – we do not calculate the standard SMT feature functions on top of weighted sentences or corpora

Cunei's Instance-Based Model

- Enables adaptation of each translation unit by scoring the **relevance** of each translation instance
- Facilitates the integration of per-instance information
- Equivalent to the standard SMT model when instance-based features are not used

Cunei's Instance-Based Model

Outperforms Moses in Czech-English and German-English

- Gain of 1.52 BLEU (6.00%) on German-English Europarl (a scenario in which SMT usually excels)
- Gain of 5.78 BLEU (21.34%) on a more complex Czech-English multi-genre evaluation

Cunei Machine Translation Platform

Try it out for yourself by visiting

`http://www.cunei.org`

The End

Cunei Machine Translation Platform

Try it out for yourself by visiting

`http://www.cunei.org`

The End

Modeling Translation Instances

Standard Approach

The fundamental unit is a phrase-pair

Uses new information to compute a new conditional likelihood of the phrase-pair

Models translation units with a weighted combination of conditional likelihoods

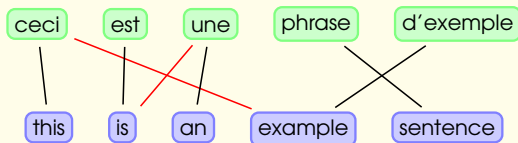
Thesis Work

The fundamental unit is an instance of translation

Uses new information to score the relevance of each translation instance

Model translation units with a weighted summation of translation instances

Alignment Sensitivity



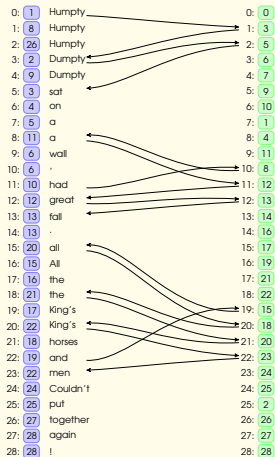
Compute likelihood by marginalizing over the alignment

$P(s t)$	$P(t s)$	$lex(s t)$	$lex(t s)$
$P(s t, d)$	$P(t s, d)$	$lex(s t, d)$	$lex(t s, d)$
$P(s t, a)$	$P(t s, a)$	$lex(s t, a)$	$lex(t s, a)$
$P(s t, d, a)$	$P(t s, d, a)$	$lex(s t, d, a)$	$lex(t s, d, a)$

Suffix Array

Humpty Dumpty sat on a wall ,
Humpty Dumpty had a great fall .
All the King's horses and all the King's men
Couldn't put Humpty together again !

Suffix Array



Locating Translation Instances

<i>POS</i>	PRP	VBZ	TO	VB	VCN	VCN	IN	DT	NNS	.
<i>Lemma</i>	it	seem	to	have	be	build	by	the	ancient	.
<i>Lexical</i>	it	seems	to	have	been	built	by	the	ancients	.

- Each type of sequence is indexed as a suffix array for efficient search
- Instances retrieved from the corpus are not required to be exact matches of the input

Generating Translation Units

- The score for each translation instance depends on the input
- Combines translation instances into $m(s_i, t_i, \lambda)$

Translation Lattice

Pruning Max 6 Pruning Ratio 0.005 Display Sequence Lexical

v	pravim	hornim	rohu	bylo	razitko	tajné	.
in	right	upper	corner	was	stamp	secret	.
in the	the	top	the corner	it was	the stamp	a secret	<null>
, in	the right	the upper	corner of	been	stamp of	the secret	up .
v	right-hand	the top		were	timestamp	classified	this .
the	the right-hand			it	stamp shall	top secret	you .
<null>	right hand			was it	rubber stamp		place .
	in the right						
	in the						
	in the right-hand						
		the upper right					
		the upper					
			upper right corner				
			corner				
			corner of				
			top right corner				
				corner		classified .	
				one corner		secret .	
				corner .			
				one corner .			

Statistical Decoder

Objective

Search the translation lattice for a set of translation units with the minimum score that completely cover the input

- Includes an inadmissible 'future cost' estimate
- Performs chart decoding to construct possible constituents, then switches to beam decoding

Second-Order Taylor Series Approximation

$$m(s_i, t_i, \lambda) = \ln \sum_{\eta} e^{\sum_k \lambda_k \cdot \phi_k(s_i, t_i, \eta)}$$

$$m(s, t, \lambda') \approx m(s, t, \lambda)$$

$$+ \sum_q (\lambda'_q - \lambda_q) \frac{\partial}{\partial \lambda_q} m(s, t, \lambda)$$

$$+ \sum_q (\lambda'_q - \lambda_q) \sum_r (\lambda'_r - \lambda_r) \frac{\partial}{\partial \lambda_q \lambda_r} m(s, t, \lambda)$$

Second-Order Taylor Series Approximation

$$\begin{aligned} m(s, t, \lambda') &\approx \ln \sum_{\eta} e^{\sum_k \lambda_k \cdot \phi_k(s, t, \eta)} \\ &\quad + \sum_q (\lambda'_q - \lambda_q) E_{\eta}[\phi_q(s, t, \eta)] \\ &\quad + \frac{1}{2} \sum_q (\lambda'_q - \lambda_q) \sum_r (\lambda'_r - \lambda_r) \\ &\quad \quad (E_{\eta}[\phi_q(s, t, \eta) \cdot \phi_r(s, t, \eta)] \\ &\quad \quad \quad - E_{\eta}[\phi_q(s, t, \eta)] \cdot E_{\eta}[\phi_r(s, t, \eta)]) \end{aligned}$$

Second-Order Taylor Series Approximation

$$\begin{aligned} m(s, t, \lambda') &\approx \ln \sum_{\eta} e^{\sum_k \lambda_k \cdot \phi_k(s, t, \eta)} \\ &+ \sum_q (\lambda'_q - \lambda_q) E_{\eta}[\phi_q(s, t, \eta)] \\ &+ \frac{1}{2} \sum_q (\lambda'_q - \lambda_q) \sum_r (\lambda'_r - \lambda_r) \\ &\quad (E_{\eta}[\phi_q(s, t, \eta) \cdot \phi_r(s, t, \eta)] \\ &\quad \quad - E_{\eta}[\phi_q(s, t, \eta)] \cdot E_{\eta}[\phi_r(s, t, \eta)]) \end{aligned}$$

Expectation used in Taylor Series

Expectation can be computed efficiently with an online update that analyzes each translation instance once

$$E_{\eta}[X] = \sum_{\eta} X \cdot P(\eta | s, t, \lambda)$$
$$P(\eta | s, t, \lambda) = \frac{e^{\sum_k \lambda_k \phi_k(s, t, \eta)}}{\sum_{\eta'} e^{\sum_k \lambda_k \phi_k(s, t, \eta')}}$$

Discounting Approximate Models

- We define a distance metric for each model approximation

$$\sum_q \left| (\lambda'_q - \lambda_q) \frac{\partial}{\partial \lambda_q} m(s, t, \lambda) \right| \\ + \sum_q \sum_r \left| (\lambda'_q - \lambda_q)(\lambda'_r - \lambda_r) \frac{\partial}{\partial \lambda_q \lambda_r} m(s, t, \lambda) \right|$$

- The log score of each (approximated) model is linearly discounted in proportion to this distance

Training Objective Function

$$(1 + e^{\mu(h) - \mu(r)}) \left(\frac{\mu(|r|)}{\mu(h)} e^{\frac{\sigma(h)}{2\mu(h)^2} - \frac{\sigma(r)}{2\mu(r)^2}} - 1 \right) \\ + \frac{\sum_{n=1}^4 \log(\mu(t_n)) - \frac{\sigma(t_n)}{2\mu(t_n)^2} - \log(\mu(c_n)) + \frac{\sigma(c_n)}{2\mu(c_n)^2}}{4}$$

m_i Log-score of hypothesis i in the n -best list

γ Gamma (used for annealing)

h Length of the hypothesis

r Length of the selected (shortest or closest) reference

c_n BLEU's "Modified count" of matching n -grams

t_n Total number of n -grams present in the hypothesis

$$p_i = \frac{e^{\gamma m_i}}{\sum_k e^{\gamma m_k}} \quad \mu(x) = \sum_i p_i x_i \quad \sigma(x) = \sum_i p_i (x_i - \mu(x))^2$$

Instance-Specific Alignment Features

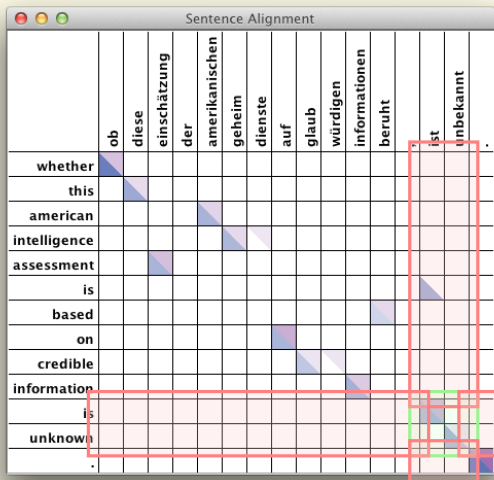
- **Inside** score
- Outside score
- Unknown score

Sentence Alignment

	ob	diese	einschätzung	der	amerikanischen	geheim	dienste	auf	glaub	würdigen	informationen	beruht	,	ist	unbekannt	.
whether	■															
this		■														
american					■											
intelligence						■										
assessment			■													
is																
based												■				
on								■								
credible									■							
information										■						
is																
unknown																
.																■

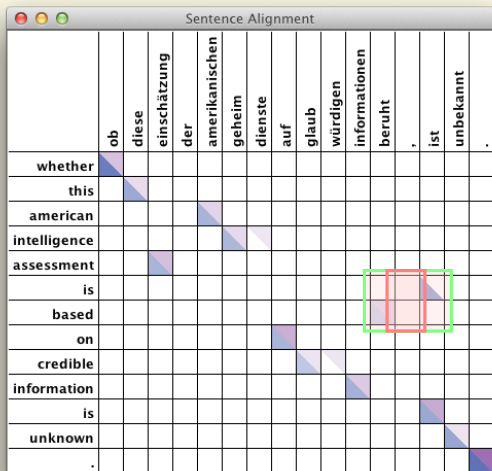
Instance-Specific Alignment Features

- Inside score
- **Outside** score
- Unknown score



Instance-Specific Alignment Features

- Inside score
- Outside score
- **Unknown** score



CzEng Test Sentence #93

Moses what with all those paper **jeřáby** ?

Cunei what with all those paper **cranes** ?

Reference what 's with all these paper cranes ?

German Europarl Test Sentence #861

<i>Moses</i>	the democratic process in côte d'ivoire is now very got off to a good start .
<i>Cunei</i>	the democratic process in côte d'ivoire is now very well .
<i>Reference</i>	the democratic process in côte d'ivoire is well under way .

CzEng Test Sentence #487

Moses driver can not **be to establish** .

Cunei driver can not **load** .

Reference the driver could not load .

CzEng Test Sentence #1347

<i>Baseline</i>	because the french use the large roman numerals , when refer to the
<i>+ Static Annotations</i>	because the french use the large roman numerals ...
<i>+ Dynamic Annotations</i>	because the french use the large roman numerals ...
<i>+ Sentence Context</i>	because the french use the large roman numerals ...
<i>+ Document Context</i>	because the french use the large roman numerals ...
<i>All Context Features</i>	because the french use capital roman numerals ...
<i>Reference</i>	since the french use capital roman numerals to refer to the

German Europarl Test Sentence #526

Baseline

i do not know exactly what the situation
in other parts of europe , in south-east
england in any event , that is a real
and **current** threat .

+ Static Annotations

... that is a real and **current** threat .

+ Dynamic Annotations

... that is a real and **current** threat .

+ Sentence Context

... that is a real and **present** threat .

+ Document Context

... that is a real and **current** threat .

+ All Context Features

... that is a real and **present** threat .

Reference

i do not know exactly the situation
across europe but in the south-east
of england this is a real and present
danger .

German Europarl Test Sentence #688

Baseline

that was the aim of the european parliament in the legislative process on clinical **review** , and i **think** that *today we can say this : this objective* has been achieved .

+ Static Annotations

... on clinical **review** , and i **think** that *today we can say this : this objective* has been achieved .

+ Dynamic Annotations

... on clinical **trials** , and i **believe** that we *can now say : this aim* has been achieved .

+ Sentence Context

... on clinical **review** , and i **think** that *today we can say this : this objective* has been achieved .

+ Document Context

... on clinical **trials** , and i **think** that *today we can say this : this objective* has been achieved .

+ All Context Features

... on clinical **trials** , and i **believe** that we *can now say : that objective* has been achieved .

Reference

this was the european parliament 's objective in the legislative procedure on clinical trials , and i believe that today we can say that this objective has been achieved .

German Europarl Test Sentence #192

Baseline

let us hope that **we in future**
, at least these guarantees can
achieve .

+ Target Context

let us hope that **in the future we**
at least , these guarantees can
achieve .

Reference

let us hope that in the future we
will at least be able to achieve
those guarantees .

CzEng Test Sentence #760

Baseline

sadi looked quizzically at garion ,
in his hands was ready for his thin
and a small knife .

+ Target Context

sadi looked quizzically at garion ,
holding ready his thin and a small
knife .

Reference

sadi looked inquiringly at garion
, holding up his slim little knife
suggestively .

German Europarl Test Sentence #5

Baseline

for some unknown reason , **appears**
my name is not included in the list
of those present .

+ Target Context

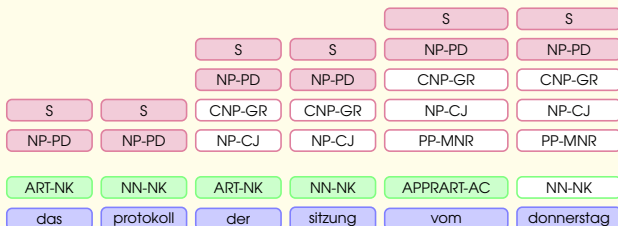
for some unknown reason , my name
is not included in the list of
those present .

Reference

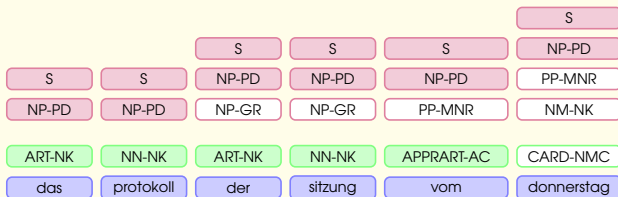
for some strange reason , my name
is missing from the register of
attendance .

Modeling Input and Replacement Similarity

Score accuracy of annotation labels



Input Phrase



Translation Instance from Corpus

German Europarl Test Sentence #363

Baseline

ultimately was after some tough negotiations , a **final outcome reached defended deserves** .

+ Annotations without Lexical Divergence

ultimately , after some tough negotiations , a **final outcome , which deserves to be defended** .

+ Annotations with Divergences

ultimately , after some tough negotiations , a **result which deserves to be defended** .

+ Annotations with Divergences and Replacement

ultimately , after some tough negotiations , a **result that deserves to be defended** .

Reference

ultimately , after some tough negotiating , an outcome was achieved that is worth defending .

German Europarl Test Sentence #255

Baseline

we all hope , of course , including
the greek colleagues here that this
dispute soon , **will now be resolved** .

+ Annotations without Lexical Divergence

... that this dispute soon **to be
resolved** .

+ Annotations with Divergences

... that this dispute soon .

+ Annotations with Divergences and Replacement

... that this dispute **will be settled**
soon .

Reference

of course we all hope - and that
includes the greek meps here - that
this dispute will soon be settled .

CzEng Test Sentence #91

Baseline

can you say to get out and podojil cow
, and i 'll do it .

*+ Annotations without
Lexical Divergence*

can you say to get out and ...

*+ Annotations with
Divergences*

can you say to get out and ...

*+ Annotations with
Divergences and
Replacement*

you can tell me to get out and ...

Reference

you can tell me to go out and milk a
cow and i 'll do it .

Static SMT-like Features

Phrase Frequency

The number of occurrences of the source phrase and the target phrase in the corpus are, respectively, c_s and c_t .

Translation.Weights.Frequency.Correlation	$\frac{(c_s - c_t)^2}{(c_s + c_t + 1)^2}$
Translation.Weights.Frequency.Source	$-\log(c_s)$
Translation.Weights.Frequency.Target	$-\log(c_t)$
Translation.Weights.Frequency.Count	$-\log(c_{s,t})$
Translation.Weights.Frequency.Counts.1	$\begin{cases} 1 & \text{if } c_{s,t} = 1 \\ 0 & \text{otherwise} \end{cases}$
Translation.Weights.Frequency.Counts.2	$\begin{cases} 1 & \text{if } c_{s,t} = 2 \\ 0 & \text{otherwise} \end{cases}$
Translation.Weights.Frequency.Counts.3	$\begin{cases} 1 & \text{if } c_{s,t} = 3 \\ 0 & \text{otherwise} \end{cases}$

Static SMT-like Features

Lexical Probability

The conditional probabilities of the source words s and target words t are relative frequency counts using the word alignments over the entire corpus.

Lexicon.Weights.Source

$$\sum_{i \in s} \max_{j \in t} \log P(s_j | t_j)$$

Lexicon.Weights.Target

$$\sum_{i \in t} \max_{j \in s} \log P(t_j | s_j)$$

Static SMT-like Features

Length Ratios

The mean, μ , and variance, σ^2 , of the lengths are calculated over the entire corpus.

Translation.Weights.Ratio.Word

$$-\frac{(|s|_{word} * \mu_{word} - |t|_{word})^2}{\sigma^2(|s|_{word} * \mu_{word} + |t|)}$$

Translation.Weights.Ratio.Character

$$-\frac{(|s|_{char} * \mu_{char} - |t|_{char})^2}{\sigma^2(|s|_{char} * \mu_{char} + |t|)}$$

Static SMT-like Features

Coverage

Let $|t|$ denote the source length of the translation unit and $|S|$ denote the length of the input sentence.

`Translation.Weights.Spans` 1

`Translation.Weights.Coverage` $\ln \frac{|t|}{|S|}$

Decoder Features

Reordering

Let the first position of the source span for the current partial translation be i and the last position of the source span for the previous partial translation be j .

`Hypothesis.Weights.Reorder.Count` $\begin{cases} 1 & \text{if } i - j \neq 1 \\ 0 & \text{otherwise} \end{cases}$

`Hypothesis.Weights.Reorder.Distance` $|i - j - 1|$

Decoder Features

Language Model

Multiple language models can be used; these refer to the model identified as `Default`. Let the order of the language model be denoted by n and the target sequence be represented as $w_0 w_1 w_2 \dots w_n$.

`LM.Default.Weights.Probability`

$$\sum_{i=0}^n \log P(w_i | w_{i-1} w_{i-2} \dots w_{i-n+1})$$

`LM.Default.Weights.Unknown`

$$\sum_{i=0}^n \begin{cases} 1 & \text{if } w_i \text{ is unknown} \\ 0 & \text{otherwise} \end{cases}$$

Decoder Features

Sentence Length

Let the phrase x contain $|x|_{word}$ words and $|x|_{char}$ characters. The mean, μ , and variance, σ^2 , of both word and character lengths are calculated over the corpus.

Sentence.Weights.Length.Words $|t|_{word}$

Sentence.Weights.Ratio.Word $-\frac{(|s|_{word} * \mu_{word} - |t|_{word})^2}{\sigma^2(|s|_{word} * \mu_{word} + |t|)}$

Sentence.Weights.Ratio.Character $-\frac{(|s|_{char} * \mu_{char} - |t|_{char})^2}{\sigma^2(|s|_{char} * \mu_{char} + |t|)}$

Phrase Alignment Features

Let $\alpha_s(i, j)$ and $\alpha_t(i, j)$ be the alignment score between the source word at position i and target word at position j (from the external word aligner).

Outside Probability

Let the set of positions in the source phrase and target phrase that are outside the phrase alignment be, respectively, s_{out} and t_{out} .

$$\text{Alignment.Outside.Source.Probability} \quad \sum_{i \in s_{out}} \log \frac{\epsilon + \sum_{j \in t_{out}} \alpha_t(i, j)}{\epsilon + \sum_j \alpha_t(i, j)}$$

$$\text{Alignment.Outside.Target.Probability} \quad \sum_{j \in t_{out}} \log \frac{\epsilon + \sum_{i \in s_{out}} \alpha_s(i, j)}{\epsilon + \sum_i \alpha_s(i, j)}$$

Phrase Alignment Features

Let $\alpha_s(i, j)$ and $\alpha_t(i, j)$ be the alignment score between the source word at position i and target word at position j (from the external word aligner).

Inside Probability

Let the set of positions in the source phrase and target phrase that are inside the phrase alignment be, respectively, s_{in} and t_{in} .

$$\text{Alignment.Inside.Source.Probability} \quad \sum_{i \in s_{in}} \log \frac{\epsilon + \sum_{j \in t_{in}} \alpha_t(i, j)}{\epsilon + \sum_j \alpha_t(i, j)}$$

$$\text{Alignment.Inside.Target.Probability} \quad \sum_{j \in t_{in}} \log \frac{\epsilon + \sum_{i \in s_{in}} \alpha_s(i, j)}{\epsilon + \sum_i \alpha_s(i, j)}$$

Phrase Alignment Features

Let $\alpha_s(i, j)$ and $\alpha_t(i, j)$ be the alignment score between the source word at position i and target word at position j (from the external word aligner).

Inside Unknown

The user-defined threshold θ identifies the value below which an alignment score is considered uncertain.

$$\text{Alignment.Inside.Source.Unknown} \quad \sum_{i \in s_{in}} \max\left(0, \frac{\theta - (\epsilon + \sum_j \alpha_t(i, j))}{\theta}\right)$$

$$\text{Alignment.Inside.Target.Unknown} \quad \sum_{j \in t_{in}} \max\left(0, \frac{\theta - (\epsilon + \sum_i \alpha_s(i, j))}{\theta}\right)$$

Source Context Features

Let A be the set annotations from the corpus that correspond to the translation instance and A' be the set of annotations for input. We will use A_X to represent the subset of annotations in A of type X . The features below are limited to the annotation types `Genre` and `Year`, but these features will be created for all annotations known to the system.

Static Mixture-Model

<code>Corpus.Sentence.Group.Web.Match</code>	$\begin{cases} 1 & \exists a \in A_{\text{Genre}} : a = \text{Web} \\ 0 & \text{otherwise} \end{cases}$
<code>Corpus.Sentence.Group.News.Match</code>	$\begin{cases} 1 & \exists a \in A_{\text{Genre}} : a = \text{News} \\ 0 & \text{otherwise} \end{cases}$
<code>Corpus.Sentence.Group.1999.Match</code>	$\begin{cases} 1 & \exists a \in A_{\text{Year}} : a = 1999 \\ 0 & \text{otherwise} \end{cases}$

Source Context Features

Let A be the set annotations from the corpus that correspond to the translation instance and A' be the set of annotations for input. We will use A_X to represent the subset of annotations in A of type X . The features below are limited to the annotation types `Genre` and `Year`, but these features will be created for all annotations known to the system.

Dynamic Comparison to Input

`Match.Divergence.Genre`

$$\ln \frac{1 + |A_{\text{Genre}} \cap A'_{\text{Genre}}|}{1 + |A_{\text{Genre}} \cup A'_{\text{Genre}}|}$$

`Match.Divergence.Year`

$$\ln \frac{1 + |A_{\text{Year}} \cap A'_{\text{Year}}|}{1 + |A_{\text{Year}} \cup A'_{\text{Year}}|}$$

Source Context Features

Left Intra-Sentential Context

Let the longest match be from position p_s to position p_e and the current translation instance being scored cover the span starting at m_s and ending at m_e .

$$\text{Match.Context.Left.1-gram} \quad \begin{cases} m_e - m_s & \text{if } m_s - p_s \geq 1 \\ m_e - m_s - 1 & \text{otherwise} \end{cases}$$

$$\text{Match.Context.Left.2-gram} \quad \begin{cases} m_e - m_s & \text{if } m_s - p_s \geq 2 \\ m_e - m_s - 1 & \text{if } m_s - p_s = 1 \\ m_e - m_s - 2 & \text{otherwise} \end{cases}$$

$$\text{Match.Context.Left.Length} \quad \sum_{i=1}^{m_e - m_s} \ln(i + m_s - p_s)$$

Source Context Features

Right Intra-Sentential Context

Let the longest match be from position p_s to position p_e and the current translation instance being scored cover the span starting at m_s and ending at m_e .

$$\text{Match.Context.Right.1-gram} \quad \begin{cases} m_e - m_s & \text{if } p_e - m_e \geq 1 \\ m_e - m_s - 1 & \text{otherwise} \end{cases}$$

$$\text{Match.Context.Right.2-gram} \quad \begin{cases} m_e - m_s & \text{if } p_e - m_e \geq 2 \\ m_e - m_s - 1 & \text{if } p_e - m_e = 1 \\ m_e - m_s - 2 & \text{otherwise} \end{cases}$$

$$\text{Match.Context.Right.Length} \quad \sum_{i=1}^{m_e - m_s} \ln(i + p_e - m_e)$$

Source Context Features

Document Context

Let $TF(t, d)$ be the count of type t in either the corpus document d or the input document d' . Let DF be the total number of documents and $DF(t)$ be the count of documents (over both the corpus and input) that contain the type t . Multiple context groups can be used; these refer to the group `Docs`.

$$\alpha_i = TF(t_i, d) \ln\left(\frac{DF + 1}{DF(t_i)}\right)$$

$$\beta_i = TF(t_i, d') \ln\left(\frac{DF + 1}{DF(t_i)}\right)$$

`Context.Group.Docs.Cosine`

$$-\ln\left(1 - \frac{\sum_i \alpha_i \beta_i}{\sqrt{\sum_i \alpha_i^2} \sqrt{\sum_i \beta_i^2}}\right)$$

`Context.Group.Docs.JensenShannon`

$$-\ln \sum_i \frac{\alpha_i \log_2 \frac{2\alpha_i}{\alpha_i + \beta_i}}{2 \sum_j \alpha_j} + \frac{\beta_i \log_2 \frac{2\beta_i}{\alpha_i + \beta_i}}{2 \sum_j \beta_j}$$

`Context.Group.Docs.Precision`

$$-\ln\left(1 - \frac{1 + \sum_i \min(\alpha_i, \beta_i)}{1 + \sum_i \beta_i}\right)$$

`Context.Group.Docs.Recall`

$$-\ln\left(1 - \frac{1 + \sum_i \min(\alpha_i, \beta_i)}{1 + \sum_i \alpha_i}\right)$$

Target Context Features

Intra-Sentential Context

Let n represent the 3-gram from the corpus that precedes the translation instance and h be the target hypothesis prior to being joined with the translation instance.

$$\text{Hypothesis.Weights.Context.1-gram} \quad \begin{cases} -1 & \text{if } n_3 \neq h_{|h|} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Hypothesis.Weights.Context.2-gram} \quad \begin{cases} -1 & \text{if } n_2 n_3 \neq h_{|h|-1} h_{|h|} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Hypothesis.Weights.Context.3-gram} \quad \begin{cases} -1 & \text{if } n_1 \dots n_3 \neq h_{|h|-2} \dots h_{|h|} \\ 0 & \text{otherwise} \end{cases}$$

Annotation Similarity Features

Sequential Annotations

Let the phrase contain n tokens. Multiple sequential annotations can be modeled simultaneously—these refer to the POS annotation type.

$$\delta(i) = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ tokens are equal} \\ 0 & \text{otherwise} \end{cases}$$

Match.Weights.POS.Divergence

$$\frac{1 + \sum_{i=0}^n \delta(i)}{1+n}$$

Annotation Similarity Features

Hierarchical Annotations

Let A be the set annotations from the corpus that correspond to the translation instance and A' be the set of annotations for input. We will use $A_X(i)$ to represent the subset of annotations in A of type X at position i . Multiple hierarchical annotations can be modeled simultaneously—these refer to the `Parse` annotation type.

`Match.Weights.Parse.Divergence`

$$\sqrt[n]{\prod_{i=0}^n \frac{1 + |A_{\text{Parse}}(i) \cap A'_{\text{Parse}}(i)|}{1 + |A_{\text{Parse}}(i) \cup A'_{\text{Parse}}(i)|}}$$



Foster, G. and Kuhn, R. (2007).

Mixture-model adaptation for SMT.

In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic. Association for Computational Linguistics.



Hildebrand, A. S., Eck, M., Vogel, S., and Waibel, A. (2005).

Adaptation of the translation model for statistical machine translation based on information retrieval.

In *Proceedings of the Tenth Annual Conference of the European Association for Machine Translation*, pages 133–142, Budapest, Hungary.



Lu, Y., Huang, J., and Liu, Q. (2007).

Improving statistical machine translation performance by training data selection and optimization.

In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 343–350, Prague, Czech Republic.



Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation.

In *2009 Conference on Empirical Methods in Natural Language Processing*, pages 708–717, Suntec, Singapore.



Och, F. J. (1999).

An efficient method for determining bilingual word classes.

In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 71–76, Bergen, Norway.



Shah, K., Barrault, L., and Schwenk, H. (2010).
Translation model adaptation by resampling.
In *Proceedings of the Joint Fifth Workshop on
Statistical Machine Translation and MetricsMATR*,
pages 392–399, Uppsala, Sweden. Association for
Computational Linguistics.



Smith, D. A. and Eisner, J. (2006).
Minimum risk annealing for training log-linear models.
In *Proceedings of the 21st International Conference
on Computational Linguistics and 44th Annual
Meeting of the Association for Computational
Linguistics*, pages 787–794, Sydney, Australia.



Vogel, S. (2005).
PESA: Phrase pair extraction as sentence splitting.
In *Machine Translation Summit X Proceedings*, pages
251–258, Phuket, Thailand.