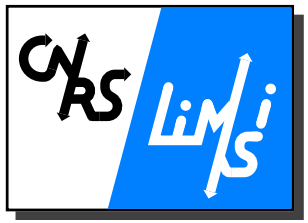
Discriminative Adaptation of Continuous Space Translation Models

Quoc-Khanh Do, Alexandre Allauzen, François Yvon

Univ. Paris-Sud / LIMSI-CNRS

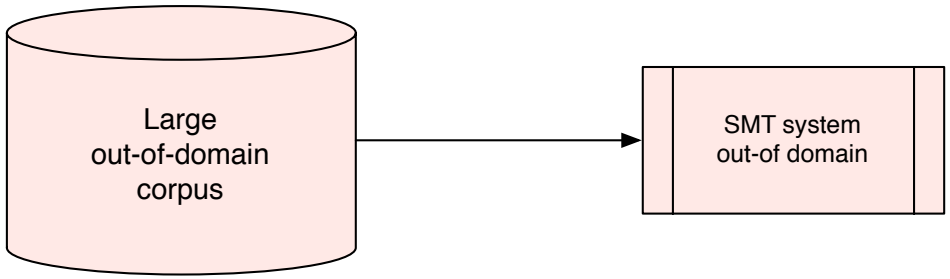


Outline

Outline

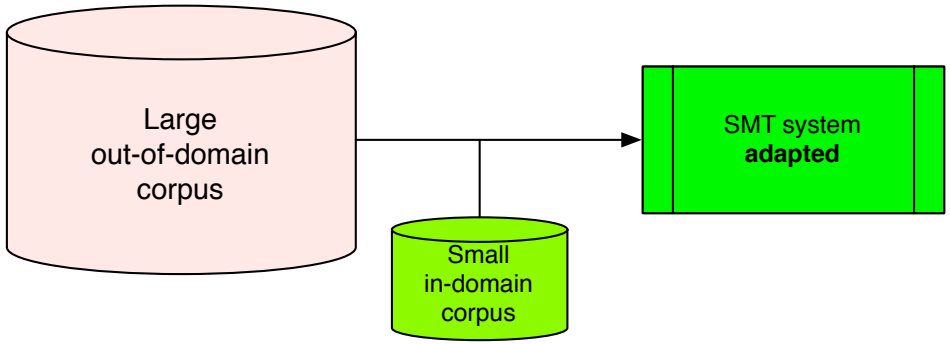
Domain adaptation in MT

or how to avoid the dilution



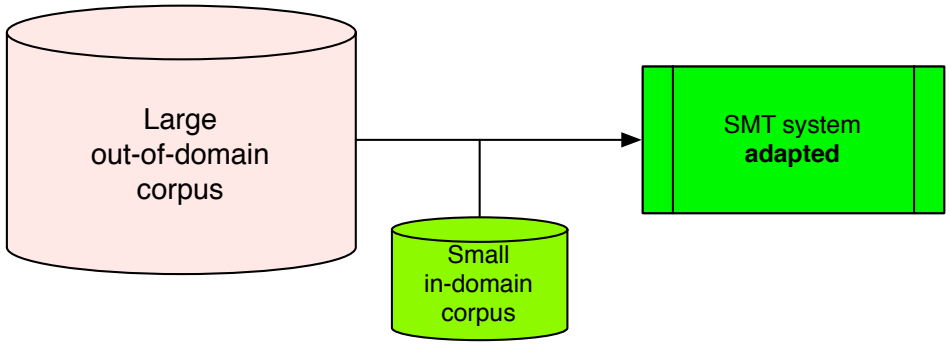
Domain adaptation in MT

or how to avoid the dilution



Domain adaptation in MT

or how to avoid the dilution



Data weighting or adaptation

See (Foster and Kuhn2007; Bertoldi and Federico2009; Axelrod et al.2011; Sennrich2012; Chen et al.2013)

- Require to retrain entirely all the models
- Very time consuming
- Especially when continuous models are involved

Continuous *vs* discrete models

Conventional discrete models

Units (words, phrases, ...) are events of discrete random variables.

- ⇒ Estimates based on relative frequencies
- ⇒ Very sparse problem
- ⇒ Ignores morphological, syntactic and semantic relationships
- ⇒ hinder the generalization power of statistical models and reduces their ability to adapt to other domains.

Continuous *vs* discrete models

Conventional discrete models

Units (words, phrases, ...) are events of discrete random variables.

⇒ Estimates based on relative frequencies

⇒ Very sparse problem

⇒ Ignores morphological, syntactic and semantic relationships

⇒ hinder the generalization power of statistical models and reduces their ability to adapt to other domains.

Continuous models

- Manipulate numerical representations of linguistic units
- Automatically trained from large corpora
- Implicitly capture some similarity relationships

⇒ A more promising power of generalization and adaptation

Discriminative Adaptation of Continuous Space Translation Models

A practical situation

- A large scale, state-of-the-art SMT system is available
- and needs to be ported to a new domain,
- using a small in-domain parallel corpus.

Our contributions

New loss functions for discriminative adaption of the CSTMs inspired from the following approaches:

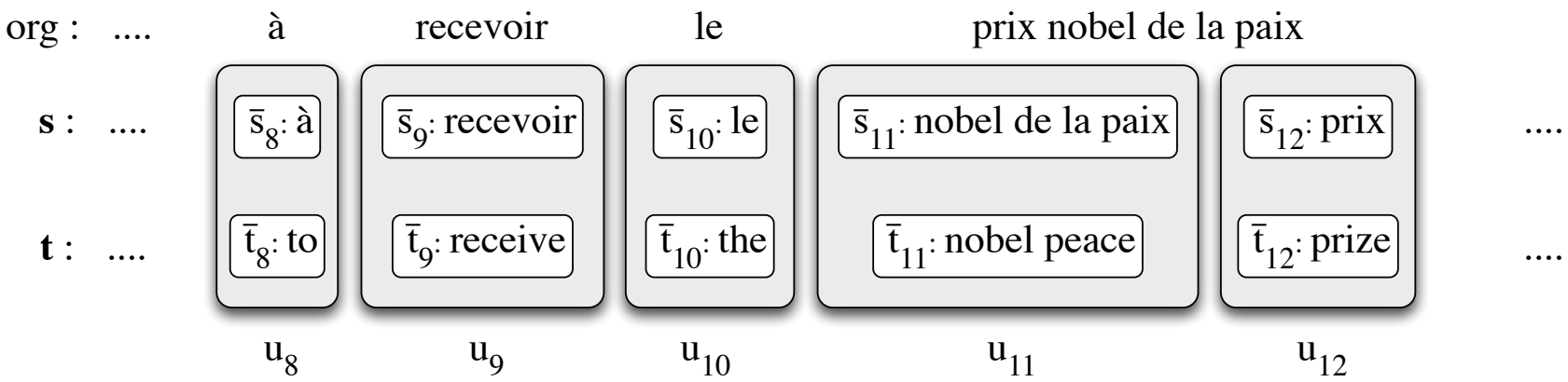
- Max-margin (Watanabe et al.2007; Cherry and Foster2012)
- Pair-wise ranking (Hopkins and May2011; Simianer et al.2012)

Case study

From News (WMT) to lecture translation (IWSLT)

Outline

The n-gram based approach in SMT



Break up the translation process (Crego and Mariño2006)

1. Source re-ordering
2. Monotonic decoding

The translation model is a n-gram of tuples:

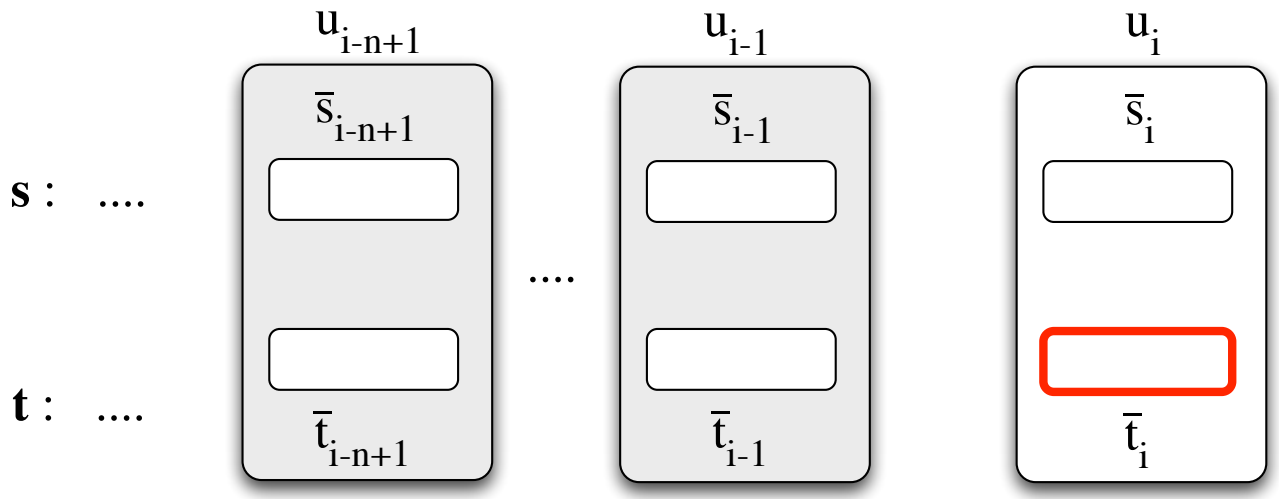
$$P(\mathbf{s}, \mathbf{t}) = \prod_{i=1}^{L} P(\mathrm{u}_i | \mathrm{u}_{i-1}, ..., \mathrm{u}_{i-n+1})$$

See `http://ncode.limsi.fr/`

A factored n-gram translation model

$$P(\mathrm{u}_i|\mathrm{u}_{i-1},...,\mathrm{u}_{i-n+1}) = P(\bar{t}_i|\bar{s}_i,\bar{s}_{i-1},\bar{t}_{i-1},...,\bar{s}_{i-n+1},\bar{t}_{i-n+1})$$

$$\times P(\bar{s}_i|\bar{s}_{i-1},\bar{t}_{i-1},...,\bar{s}_{i-n+1},\bar{t}_{i-n+1})$$

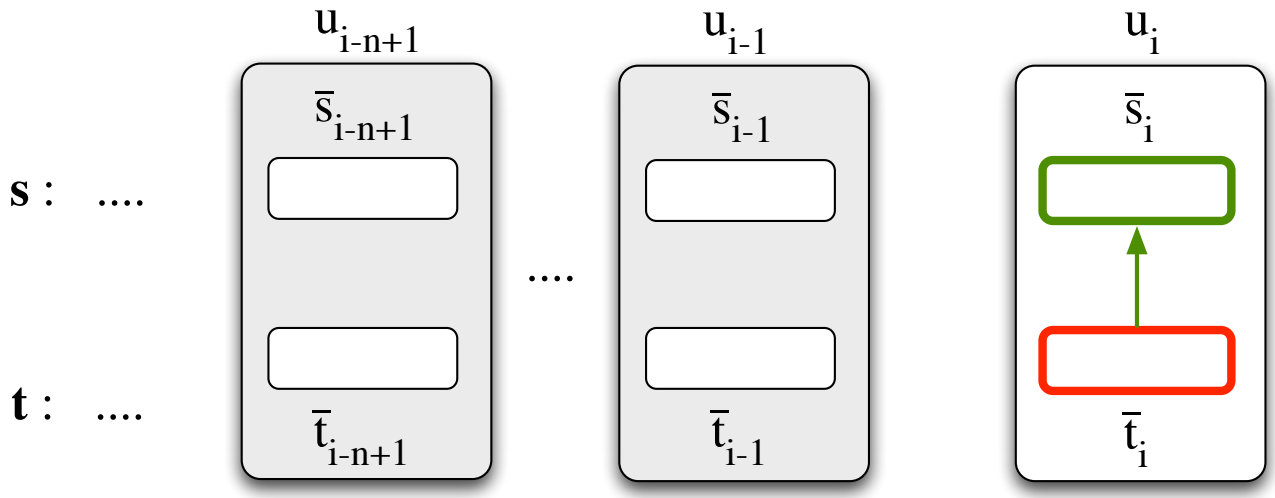


A factored n-gram translation model

$$P(\mathrm{u}_i|\mathrm{u}_{i-1},...,\mathrm{u}_{i-n+1}) = P(\overline{t}_i|\overline{s}_i,\overline{s}_{i-1},\overline{t}_{i-1},...,\overline{s}_{i-n+1},\overline{t}_{i-n+1})$$

$$\times P(\overline{s}_i|\overline{s}_{i-1},\overline{t}_{i-1},...,\overline{s}_{i-n+1},\overline{t}_{i-n+1})$$

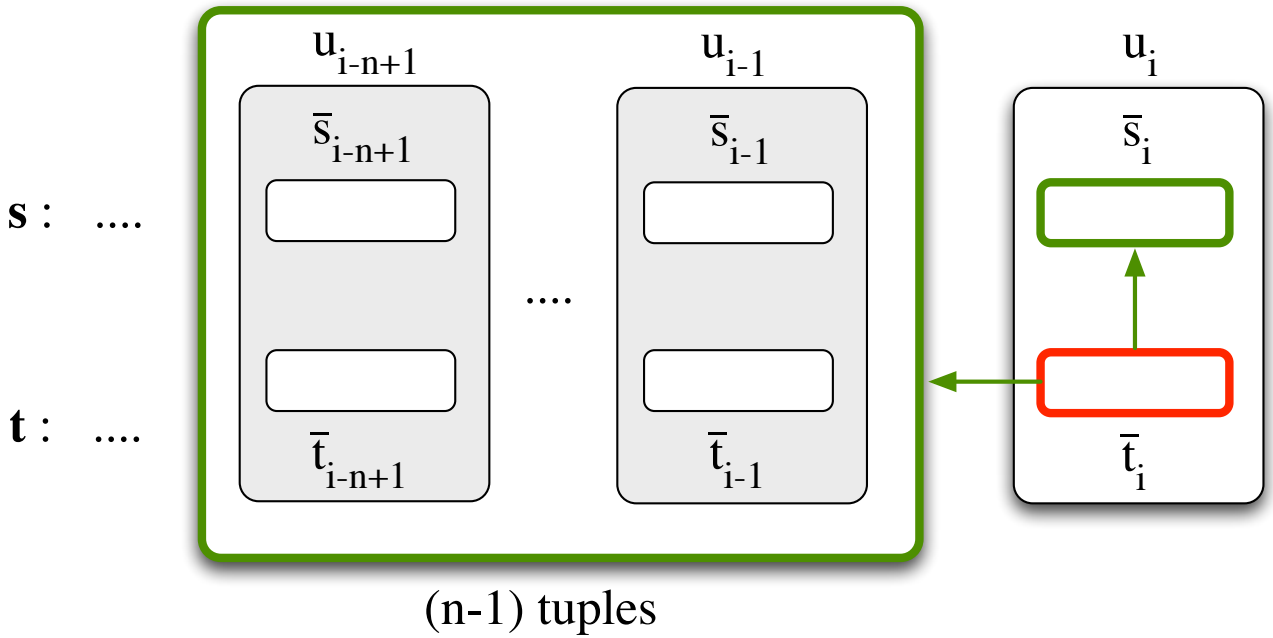
conditional translation model



A factored n-gram translation model

$$P(\mathrm{u}_i|\mathrm{u}_{i-1},...,\mathrm{u}_{i-n+1}) = P(\overline{t}_i|\overline{s}_i, \overline{s}_{i-1},\overline{t}_{i-1},...,\overline{s}_{i-n+1},\overline{t}_{i-n+1}) \times P(\overline{s}_i|\overline{s}_{i-1},\overline{t}_{i-1},...,\overline{s}_{i-n+1},\overline{t}_{i-n+1})$$

conditional translation model

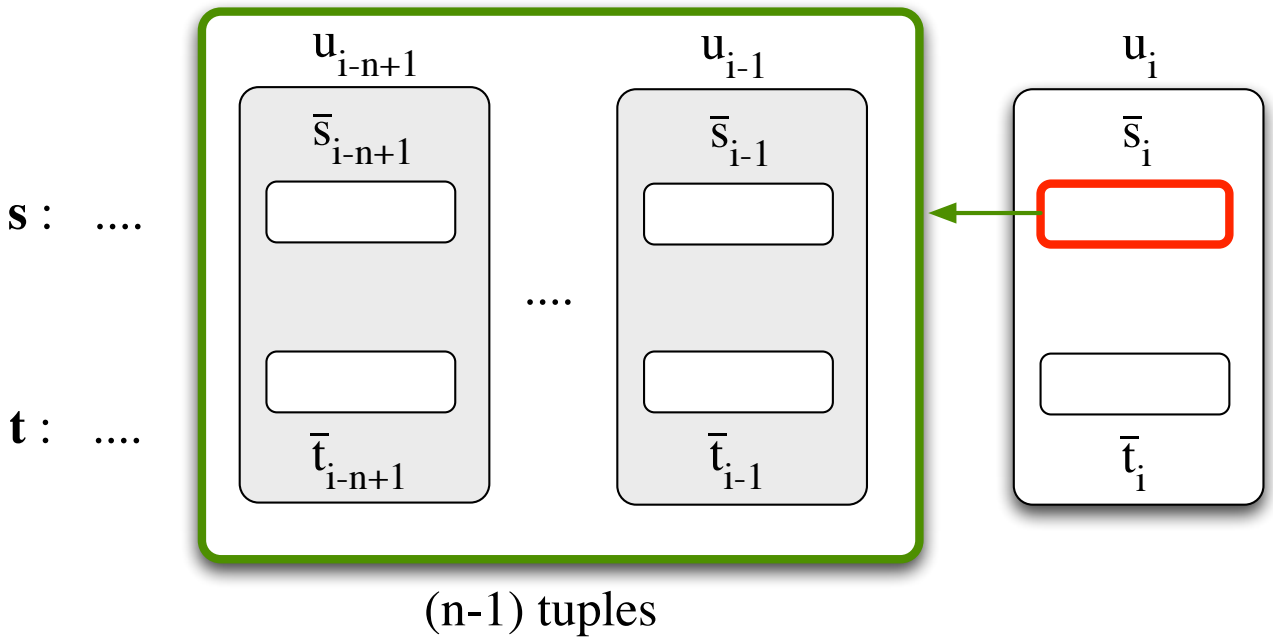


A factored n-gram translation model

$$P(\mathrm{u}_i|\mathrm{u}_{i-1},...,\mathrm{u}_{i-n+1}) = P(\bar{t}_i|\bar{s}_i,\bar{s}_{i-1},\bar{t}_{i-1},...,\bar{s}_{i-n+1},\bar{t}_{i-n+1})$$

$$\times P(\bar{s}_i \,|\, \bar{s}_{i-1},\bar{t}_{i-1},...,\bar{s}_{i-n+1},\bar{t}_{i-n+1})$$

conditional translation model «distortion» model



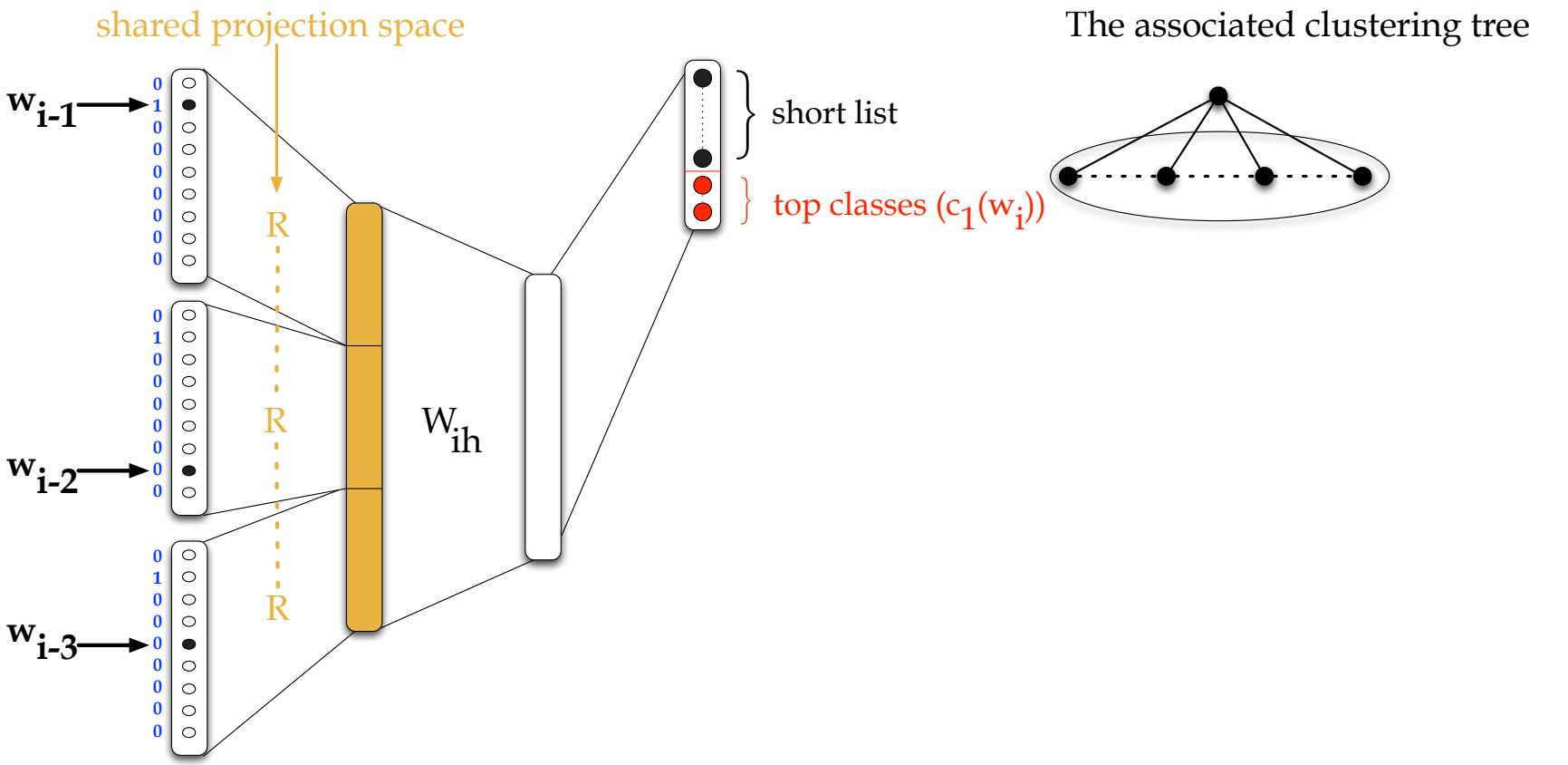
A factored n-gram translation model

$$P(\mathrm{u}_i|\mathrm{u}_{i-1},...,\mathrm{u}_{i-n+1}) = P(\bar{t}_i|\bar{s}_i,\bar{s}_{i-1},\bar{t}_{i-1},...,\bar{s}_{i-n+1},\bar{t}_{i-n+1})$$
$$\times P(\bar{s}_i|\bar{s}_{i-1},\bar{t}_{i-1},...,\bar{s}_{i-n+1},\bar{t}_{i-n+1})$$

- These distributions can be decomposed at the level of words, and
- estimated with the bilingual version of the SOUL model (Le et al.2012)

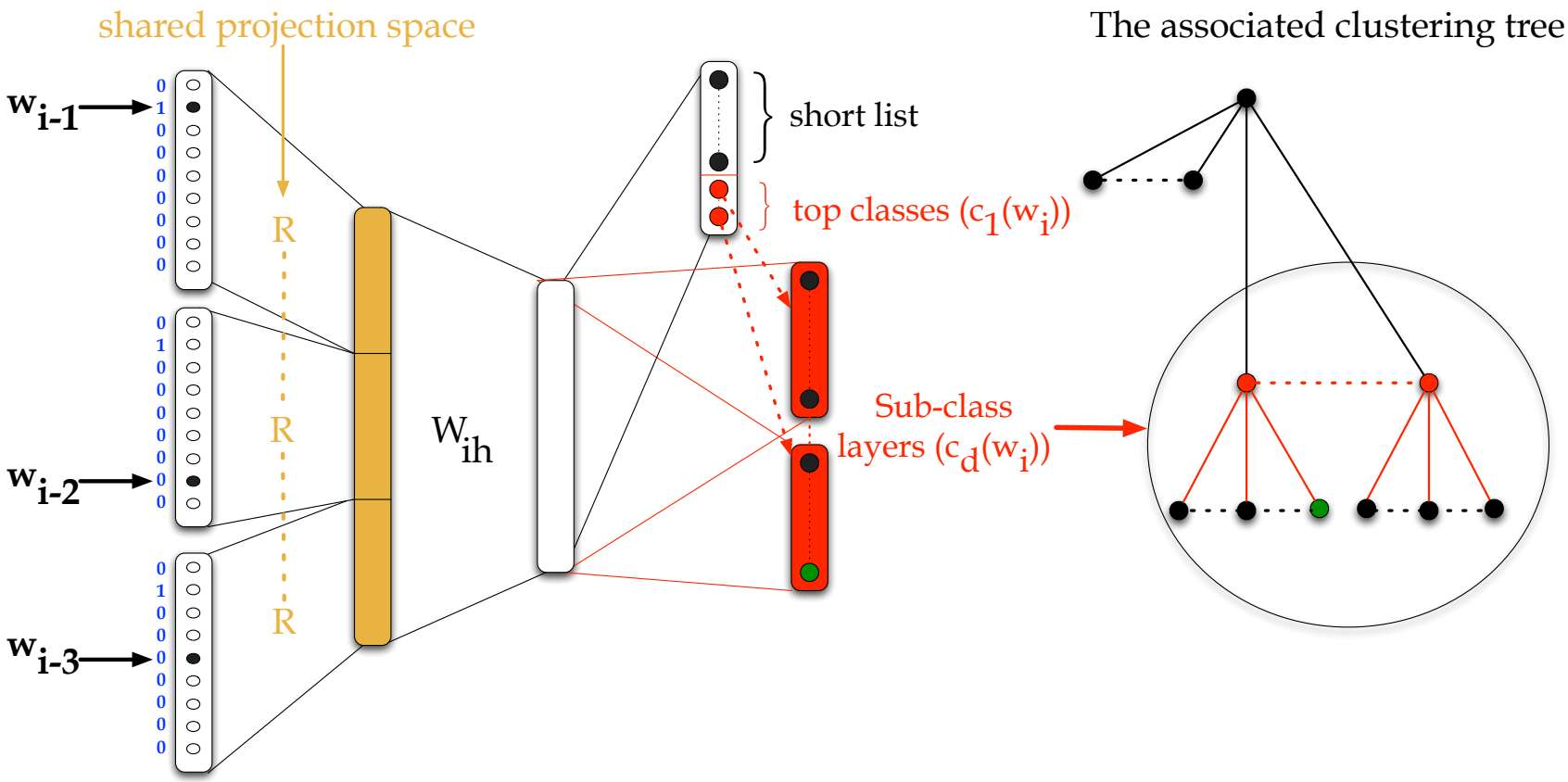
The SOUL model in one picture

$$P(w_i|h) = P(c_1(w_i)|h)$$



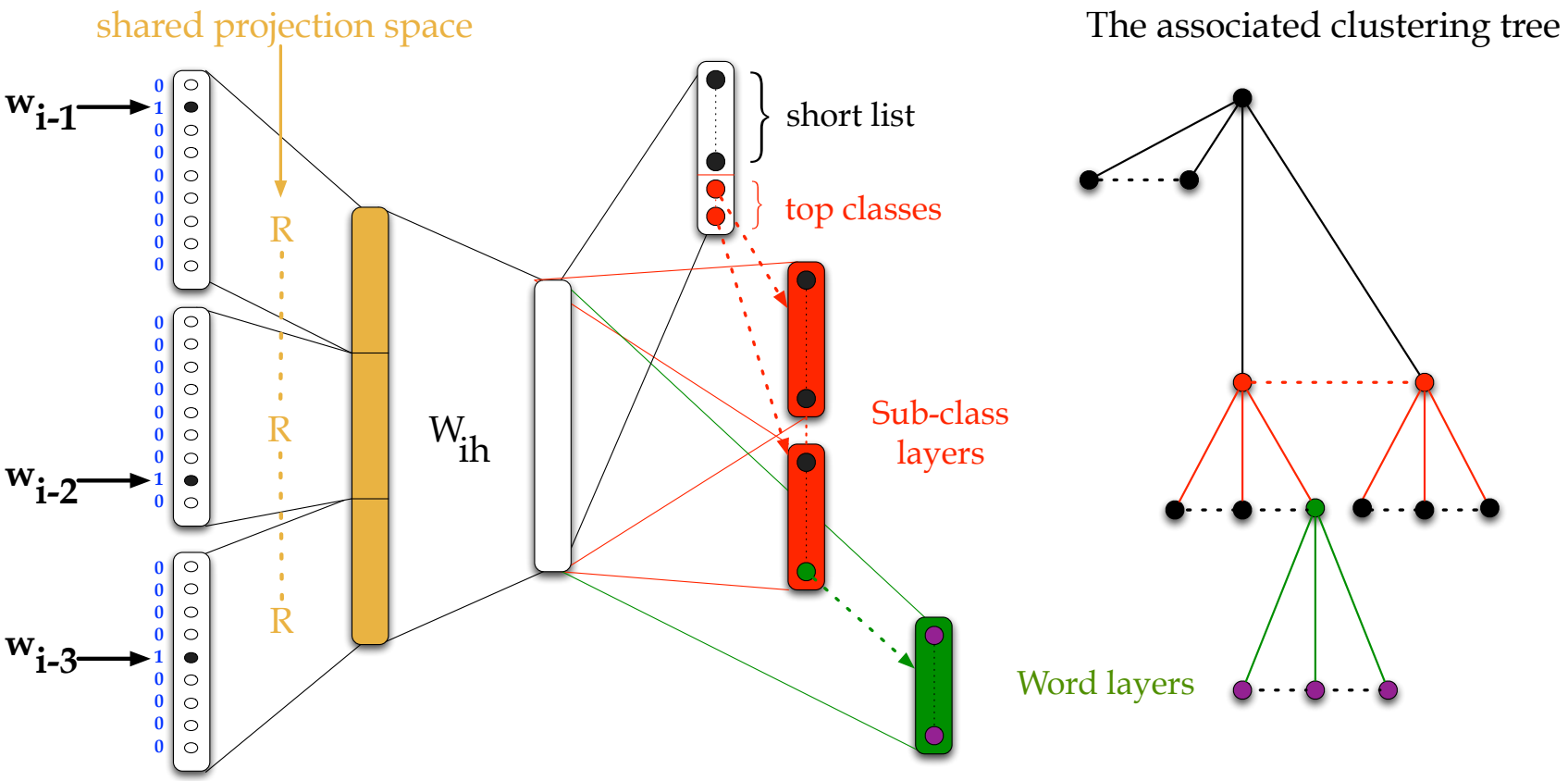
The SOUL model in one picture

$$P(w_i|h) = P(c_1(w_i)|h) \times \prod_{d=2}^{D} P(c_d(w_i)|h, c_{1:d-1}(w_i))$$



The SOUL model in one picture

$$P(w_i|h) = P(c_1(w_i)|h) \times \prod_{d=2}^{D} P(c_d(w_i)|h, c_{1:d-1}(w_i))$$



Translation modelling with SOUL

Standard word n-gram models (monolingual)

Successful in Automatic Speech Recognition and SMT (Le et al.2011; Allauzen et al.2011).

Translation modelling with SOUL

Standard word n-gram models (monolingual)

Successful in Automatic Speech Recognition and SMT (Le et al.2011; Allauzen et al.2011).

Word factored translation models (Le et al.2012)

They involve two languages:

- the predicted word is in target or source language
- the context is made of both source and target words

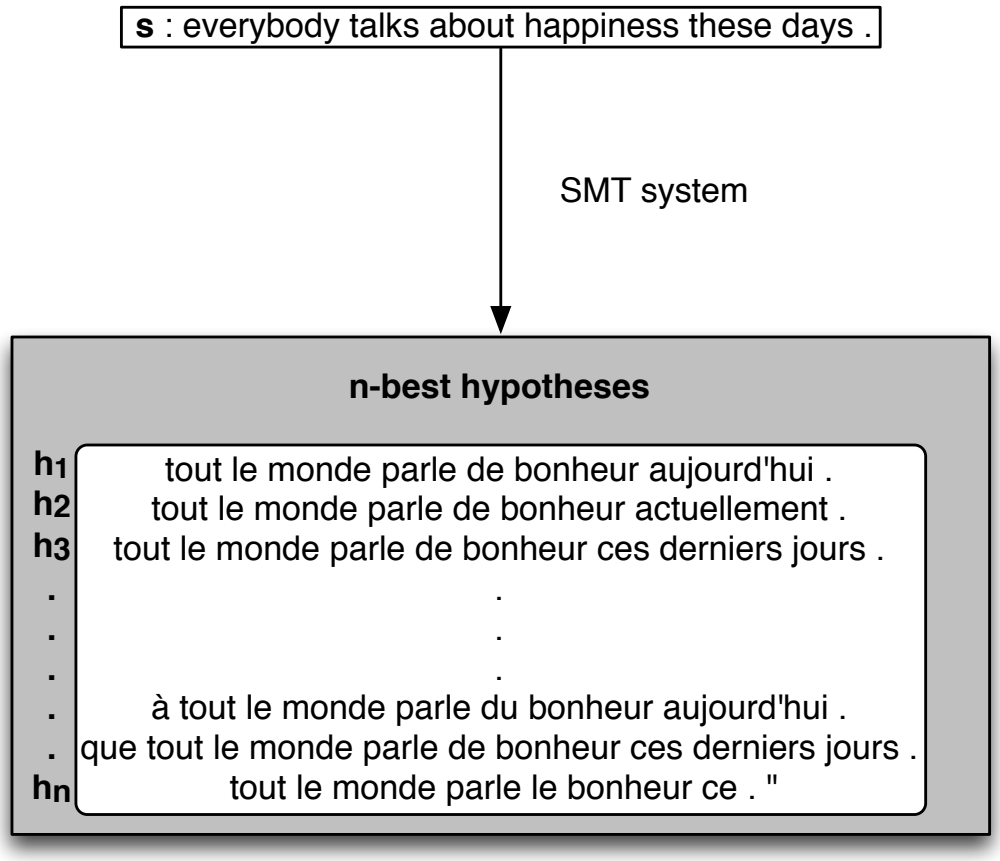
$\Rightarrow$ Two different projection matrices ($\mathbf{R}$).

Conventional training

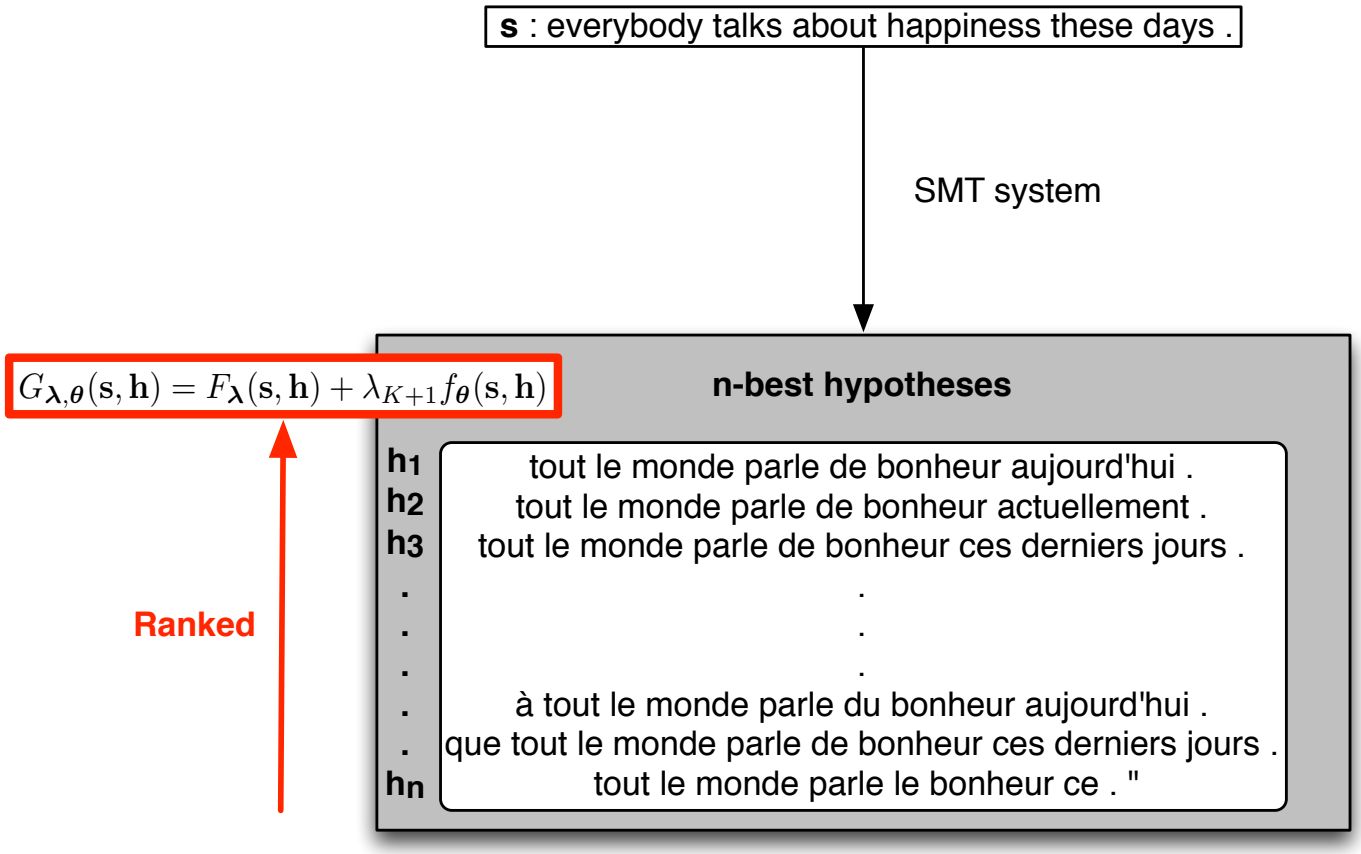
Maximize the log-likelihood of the bilingual training data

Outline

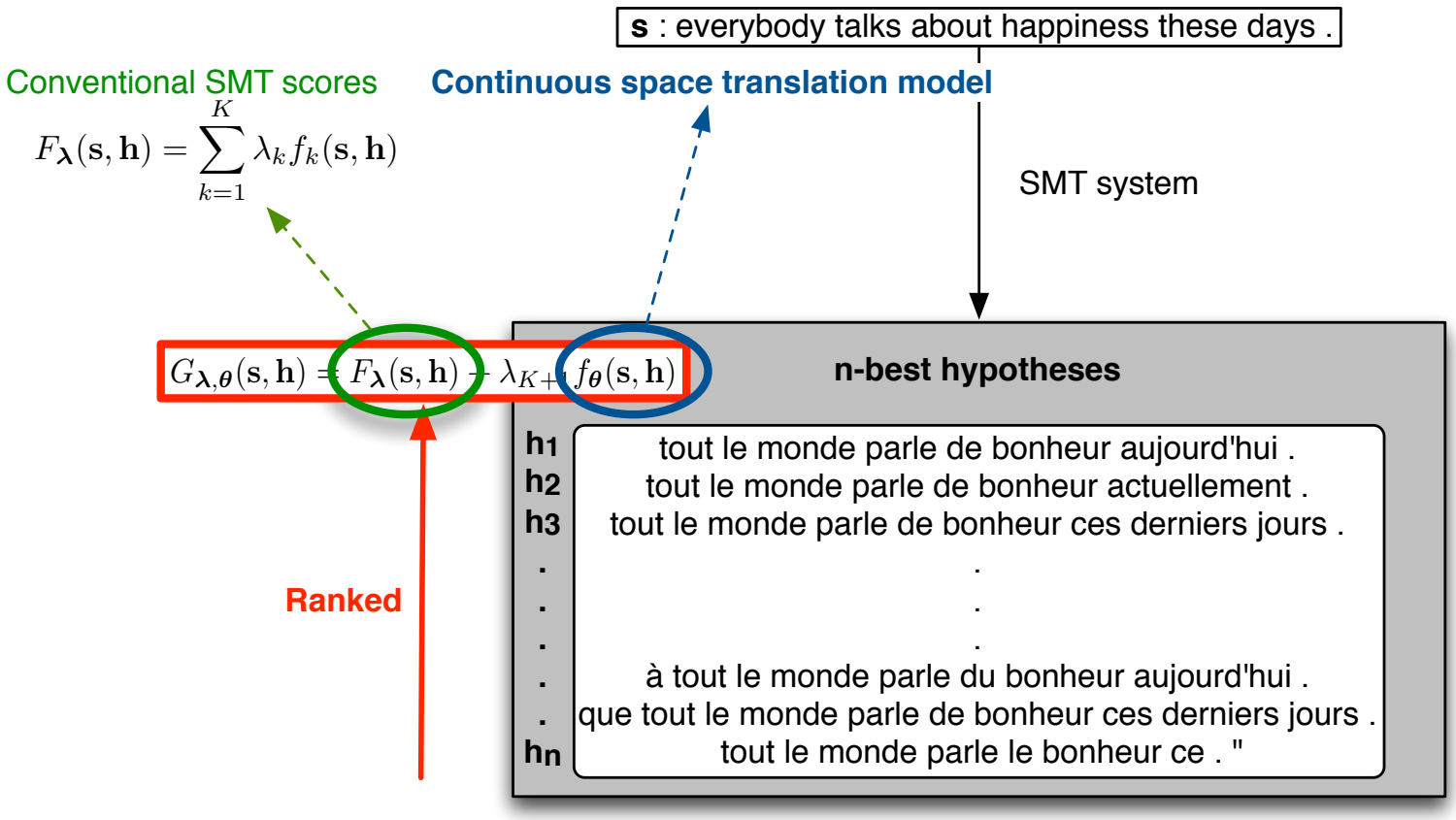
N-best-list rescoring



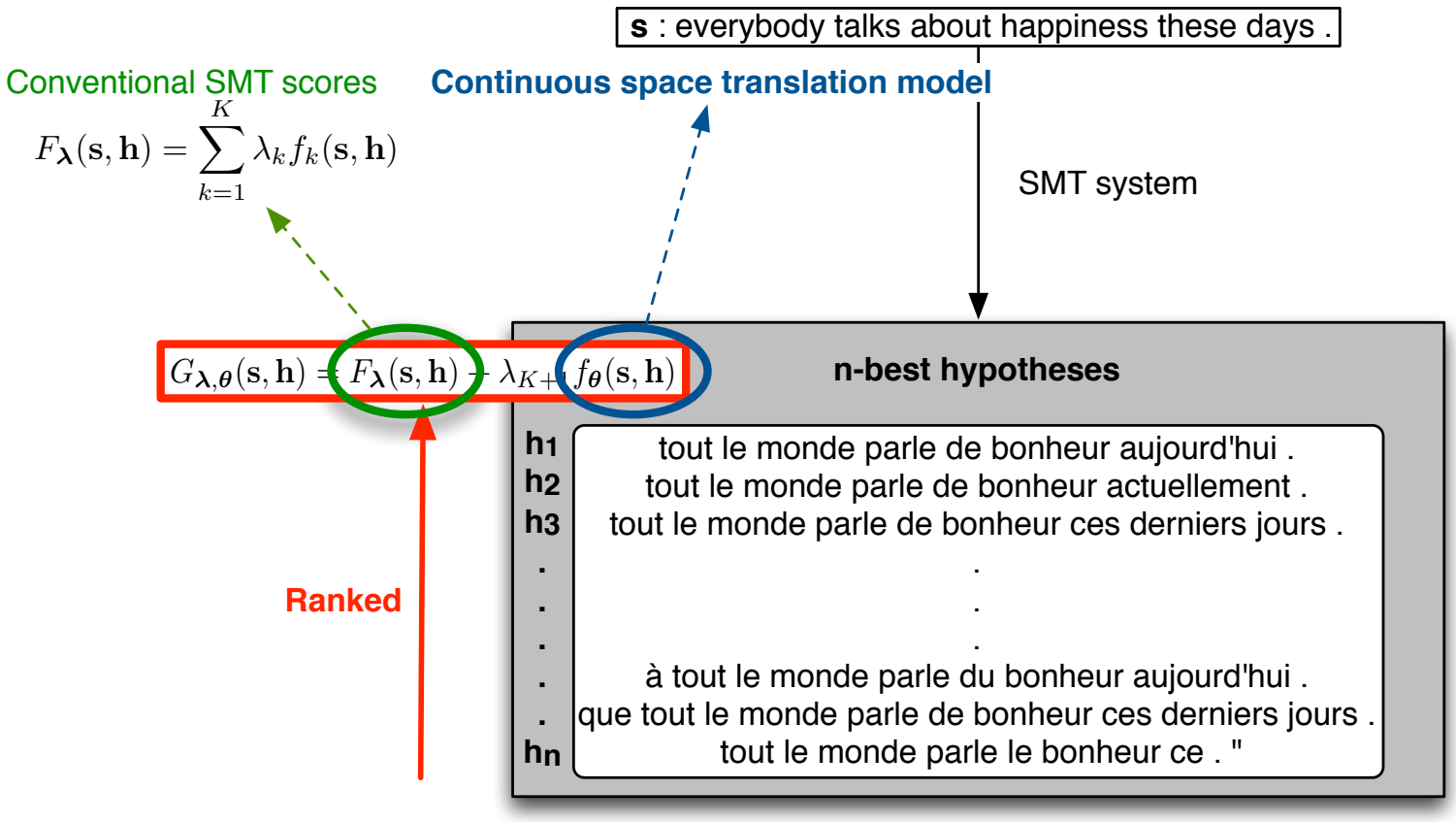
N-best-list rescoring



N-best-list rescoring



N-best-list rescoring



alternatively tune the vector of coefficients $\boldsymbol{\lambda}$
and adapt the CSTM's parameters $\boldsymbol{\theta}$

Optimization algorithm

Algorithm 1 Joint optimization procedure for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$

Initialize $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$
for each iteration **do**
    for M mini-batches **do** ▷ $\boldsymbol{\lambda}$ is fixed
        Compute the sub-gradient of $\mathcal{L}(\boldsymbol{\theta}, \mathbf{s})$ for all $\mathbf{s}$ in the mini-batch
        Update $\boldsymbol{\theta}$
    end for
    Update $\boldsymbol{\lambda}$ using dev set ▷ $\boldsymbol{\theta}$ is fixed
end for

Optimization algorithm

Algorithm 1 Joint optimization procedure for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$

1: Initialize $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$
2: **for** each iteration **do**
3: **for** M mini-batches **do** ▷ $\boldsymbol{\lambda}$ is fixed
4: Compute the sub-gradient of $\mathcal{L}(\boldsymbol{\theta}, \mathbf{s})$ for all $\mathbf{s}$ in the mini-batch
5: Update $\boldsymbol{\theta}$
6: **end for**
7: Update $\boldsymbol{\lambda}$ using dev set ▷ $\boldsymbol{\theta}$ is fixed
8: **end for**

Optimization algorithm

Algorithm 1 Joint optimization procedure for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$

1: Initialize $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$
2: **for** each iteration **do**
3: **for** M mini-batches **do** ▷ $\boldsymbol{\lambda}$ is fixed
4: Compute the sub-gradient of $\mathcal{L}(\boldsymbol{\theta}, \mathbf{s})$ for all $\mathbf{s}$ in the mini-batch
5: Update $\boldsymbol{\theta}$
6: **end for**
7: Update $\boldsymbol{\lambda}$ using dev set ▷ $\boldsymbol{\theta}$ is fixed
8: **end for**

Optimization algorithm

Algorithm 1 Joint optimization procedure for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$

1: Initialize $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$
2: **for** each iteration **do**
3: **for** M mini-batches **do** ▷ $\boldsymbol{\lambda}$ is fixed
4: Compute the sub-gradient of $\mathcal{L}(\boldsymbol{\theta}, \mathbf{s})$ for all $\mathbf{s}$ in the mini-batch
5: Update $\boldsymbol{\theta}$
6: **end for**
7: Update $\boldsymbol{\lambda}$ using dev set ▷ $\boldsymbol{\theta}$ is fixed
8: **end for**

- MERT or KBMIRA (Cherry and Foster2012)

Optimization algorithm

Algorithm 1 Joint optimization procedure for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$

1: Initialize $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$
2: **for** each iteration **do**
3: **for** M mini-batches **do** ▷ $\boldsymbol{\lambda}$ is fixed
4: Compute the sub-gradient of $\mathcal{L}(\boldsymbol{\theta}, \mathbf{s})$ for all $\mathbf{s}$ in the mini-batch
5: Update $\boldsymbol{\theta}$
6: **end for**
7: Update $\boldsymbol{\lambda}$ using dev set ▷ $\boldsymbol{\theta}$ is fixed
8: **end for**

- MERT or KBMIRA (Cherry and Foster2012)
- $\mathcal{L}(\theta, s)$???

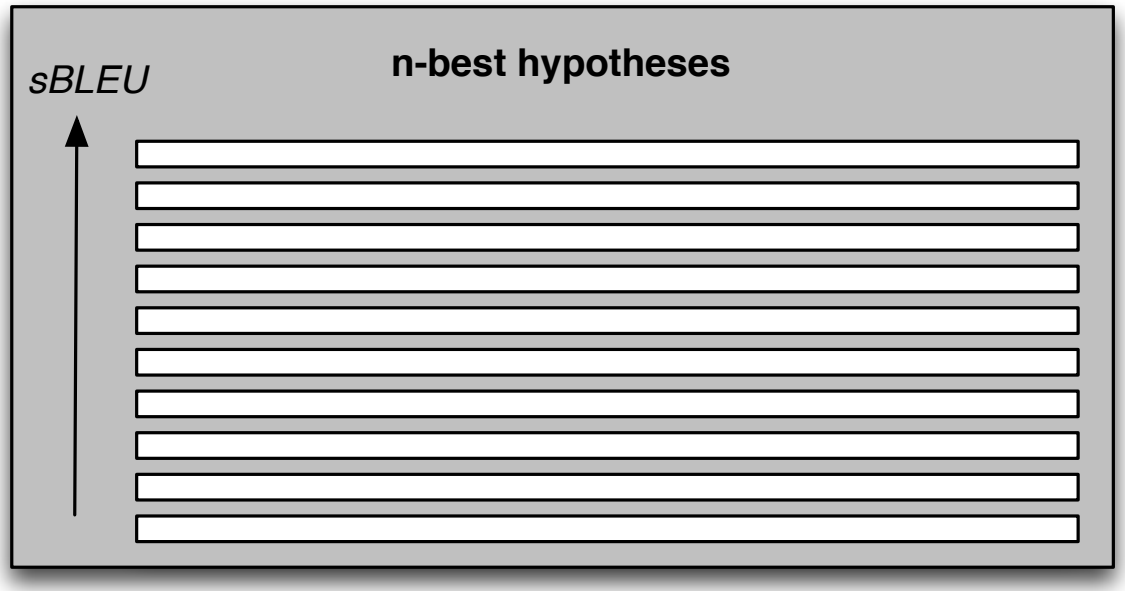
Max-margin adaptation

The cost of an hypothesis:

$$\text{cost}_\alpha(\mathbf{h}) = \alpha\big(sBLEU(\mathbf{h}^*) - sBLEU(\mathbf{h})\big), \text{ where}$$
$$\mathbf{h}^* = \underset{\mathbf{h}}{\text{argmax}}\, sBLEU(\mathbf{h}), \text{ the best hypothesis}$$

$sBLEU$ is the sentence BLEU and α a scaling factor

Max-margin loss function



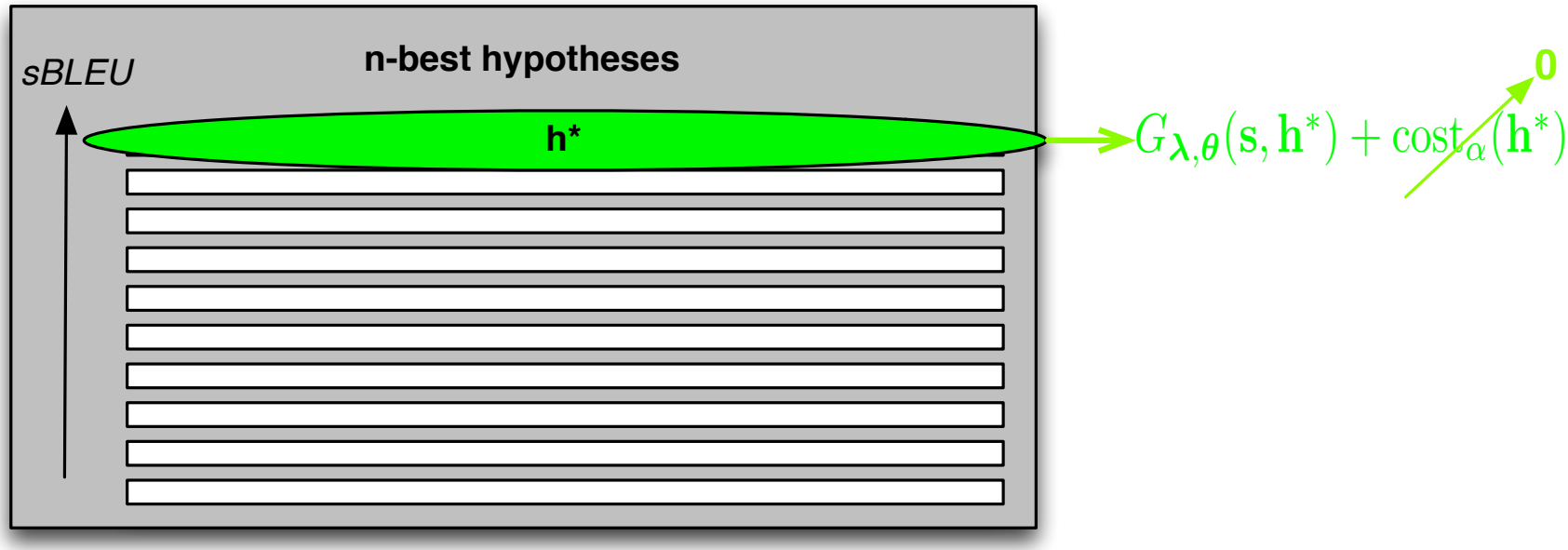
Max-margin adaptation

The cost of an hypothesis:

$$\text{cost}_\alpha(\mathbf{h}) = \alpha\big(sBLEU(\mathbf{h}^*) - sBLEU(\mathbf{h})\big), \text{ where}$$
$$\mathbf{h}^* = \underset{\mathbf{h}}{\text{argmax}}\, sBLEU(\mathbf{h}), \text{ the best hypothesis}$$

$sBLEU$ is the sentence BLEU and α a scaling factor

Max-margin loss function



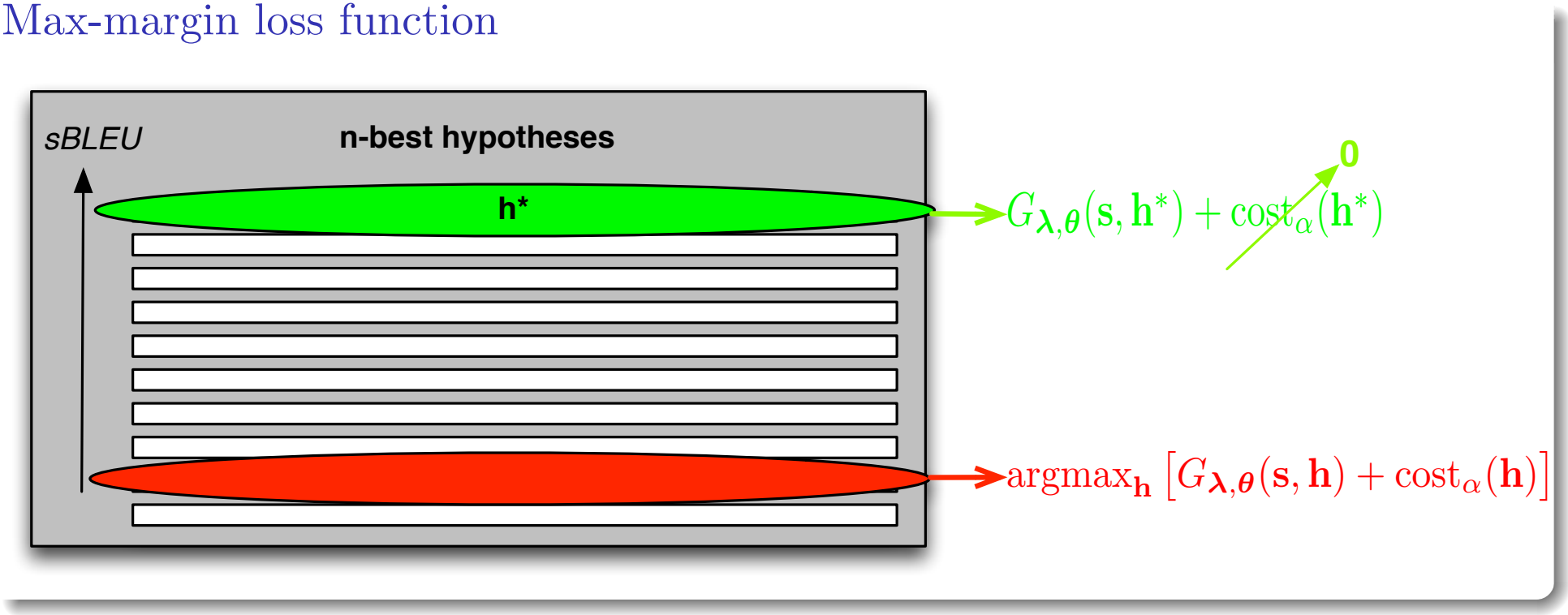
Max-margin adaptation

The cost of an hypothesis:

$$\text{cost}_\alpha(\mathbf{h}) = \alpha\big(sBLEU(\mathbf{h}^*) - sBLEU(\mathbf{h})\big), \text{ where}$$
$$\mathbf{h}^* = \underset{\mathbf{h}}{\text{argmax}}\, sBLEU(\mathbf{h}), \text{ the best hypothesis}$$

$sBLEU$ is the sentence BLEU and α a scaling factor

Max-margin loss function



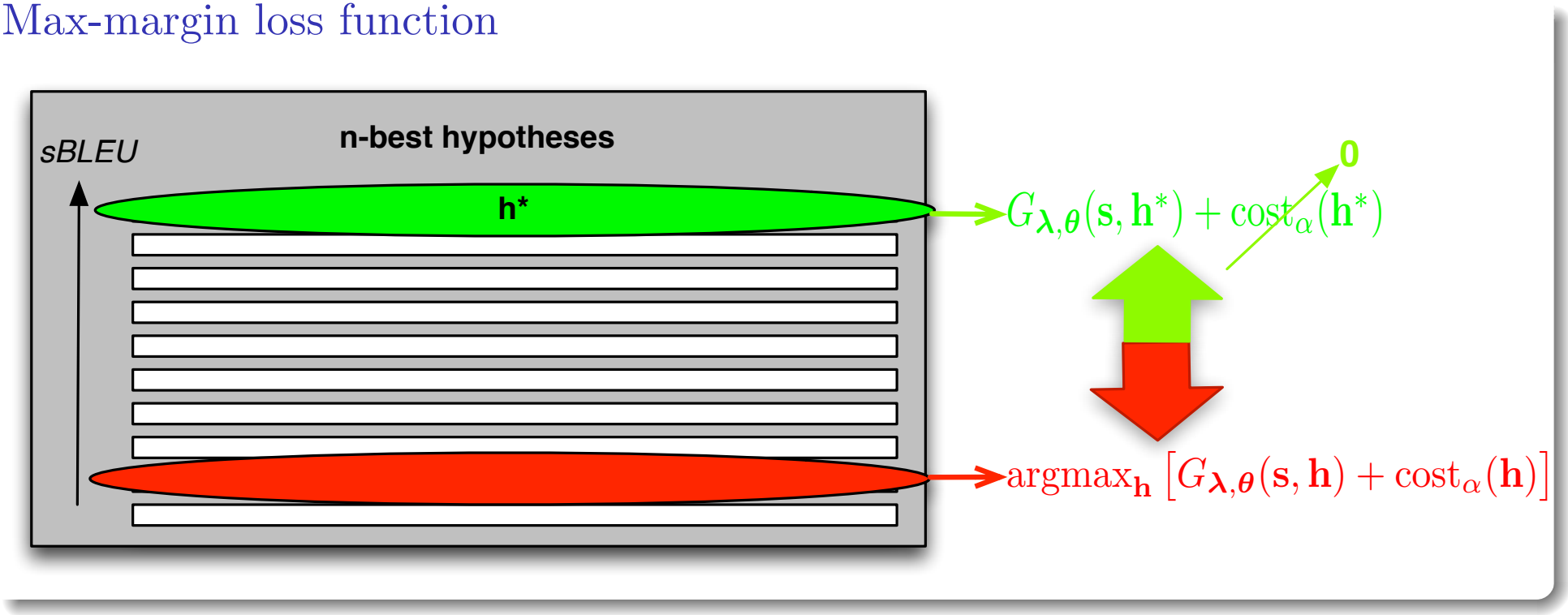
Max-margin adaptation

The cost of an hypothesis:

$$\text{cost}_\alpha(\mathbf{h}) = \alpha\big(sBLEU(\mathbf{h}^*) - sBLEU(\mathbf{h})\big), \text{ where}$$
$$\mathbf{h}^* = \underset{\mathbf{h}}{\text{argmax}}\, sBLEU(\mathbf{h}), \text{ the best hypothesis}$$

$sBLEU$ is the sentence BLEU and α a scaling factor

Max-margin loss function



Combining the Max-margin and Pairwise-ranking approaches

Drawbacks of the Max-margin approach

- Only considers a pair of hypotheses
- While several good translations only differ slightly

→ Pairwise-ranking (Hopkins and May2011; Simianer et al.2012)

→ Including the margin term (the cost)

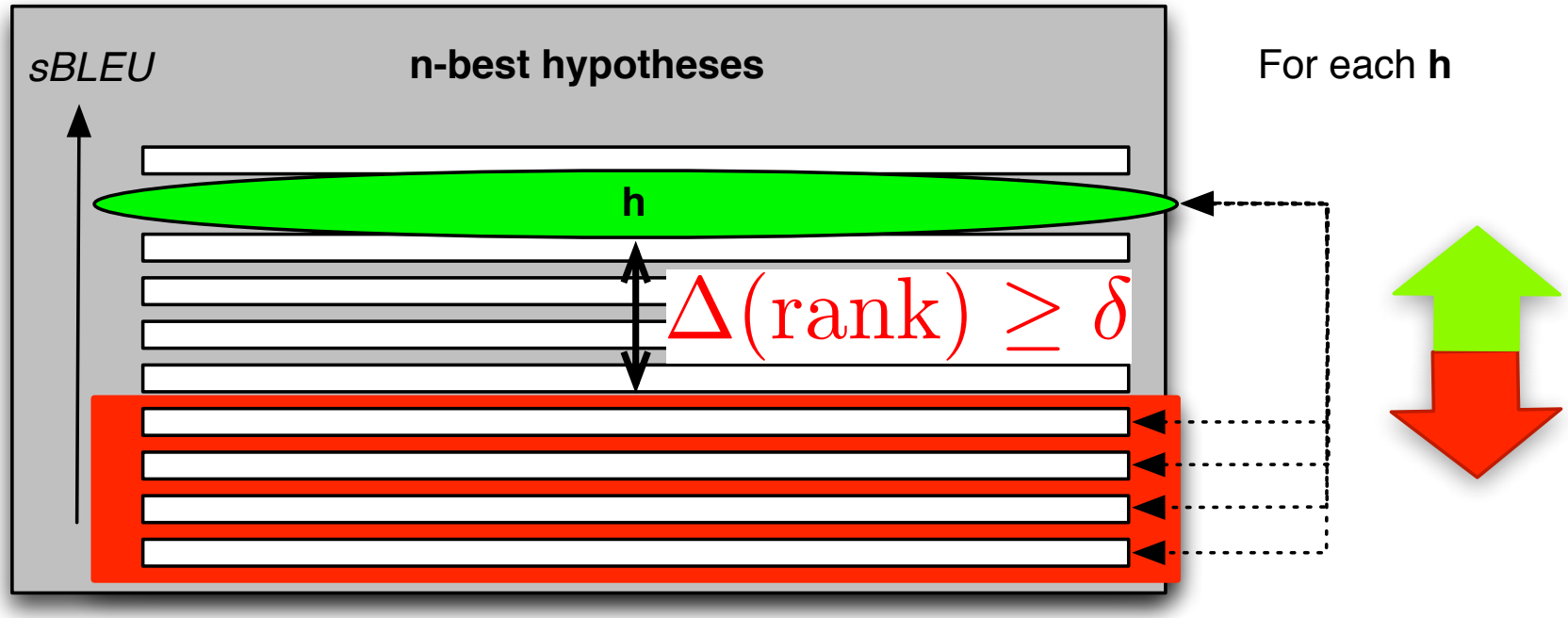
Combining the Max-margin and Pairwise-ranking approaches

Drawbacks of the Max-margin approach

- Only considers a pair of hypotheses
- While several good translations only differ slightly

→ Pairwise-ranking (Hopkins and May2011; Simianer et al.2012)

→ Including the margin term (the cost)



Outline

Experimental set-up

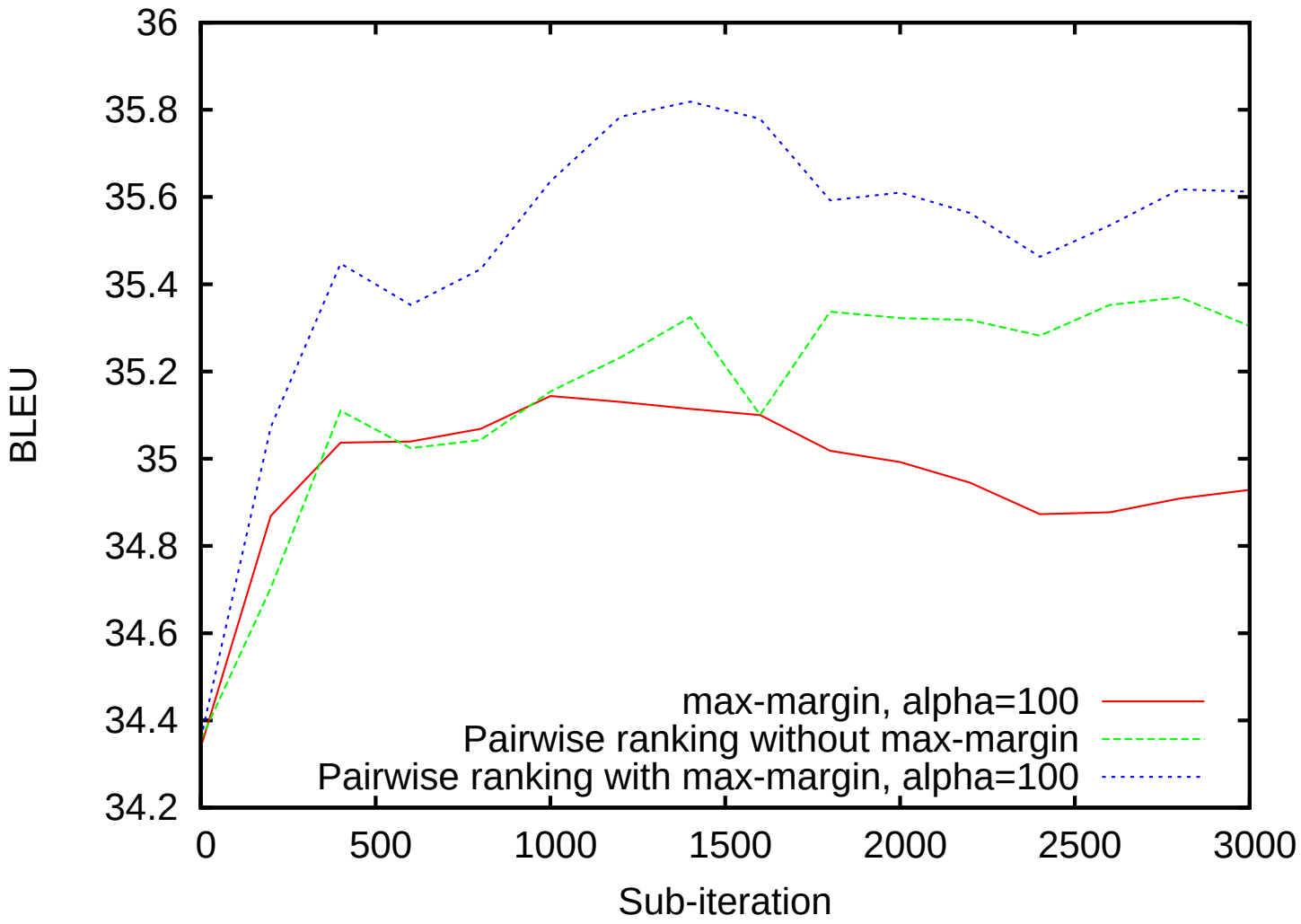
	English	French
Vocabulary	505K	492K
Out-of-domain data	12M sentences (WMT'13)	
In-domain data	107,058 sentences (IWSLT'11)	
Development set	1,664 sentences	
Test set	934 sentences	
N-best lists	300 hypotheses	
Context length CSTM	9	

Experimental set-up

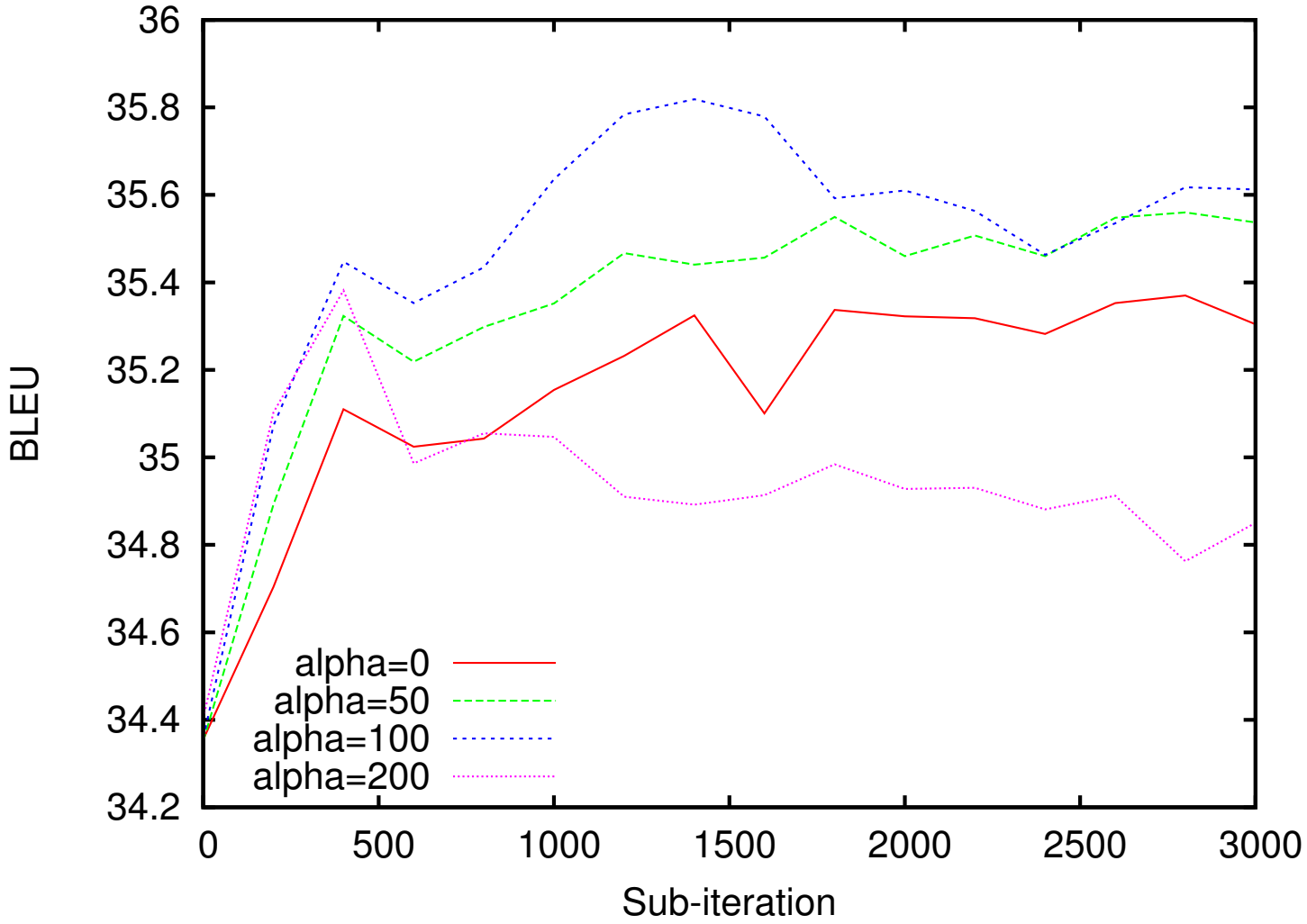
	English	French
Vocabulary	505K	492K
Out-of-domain data	12M sentences (WMT'13)	
In-domain data	107,058 sentences (IWSLT'11)	
Development set	1,664 sentences	
Test set	934 sentences	
N-best lists	300 hypotheses	
Context length CSTM	9	

- Out-of-domain data to train baseline system (N-code + CSTM)
- In-domain data to adapt CSTM to TED Talks task.

Comparison of different loss functions



Impact of the margin - α



Final results

System	*dev*	*test*
Baseline systems (out-of-domain)		
n-code	33.9	27.6
n-code + CSTM WMT	34.4	28.5
Adapted systems		
n-code + CSTM CLL adapted	35.0	29.1
n-code + CSTM $\mathcal{L}_{mm}$ adapted $\alpha = 100$	35.1	29.4
n-code + CSTM $\mathcal{L}_{pro}$ adapted	35.4	29.5
n-code + CSTM $\mathcal{L}_{pro-mm}$ adapted $\alpha = 100$	35.8	29.6
n-code + all WMT CSTMs + 2 CSTMs $\mathcal{L}_{pro-mm}$	**36.4**	**29.9**

Outline

Conclusion

- Discriminative criteria on N-best lists
- Joint optimization of CSTM and SMT system's parameters
- BLEU-based margins introduced
- Pair-wise ranking versus max-margin

Conclusion

- Discriminative criteria on N-best lists
- Joint optimization of CSTM and SMT system's parameters
- BLEU-based margins introduced
- Pair-wise ranking versus max-margin

+ Future works : Other criteria (expected-BLEU (Gao and He2013)), other model structures


Alexandre Allauzen, Hélène Bonneau-Maynard, Hai-Son Le, Aurélien Max, Guillaume Wisniewski, François Yvon, Gilles Adda, Josep Maria Crego, Adrien Lardilleux, Thomas Lavergne, and Artem Sokolov.
2011.
LIMSI @ WMT11.
In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 309–315, Edinburgh, Scotland.


Amittai Axelrod, Xiaodong He, and Jianfeng Gao.
2011.
Domain adaptation via pseudo in-domain data selection.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 355–362.


Nicola Bertoldi and Marcello Federico.
2009.
Domain adaptation for statistical machine translation with monolingual resources.
In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece.


Boxing Chen, Roland Kuhn, and George Foster.

2013.
Vector space model for adaptation in statistical machine translation.
In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1285–1293.


Colin Cherry and George Foster.
2012.
Batch tuning strategies for statistical machine translation.
In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 427–436, June.


Josep M. Crego and José B. Mariño.
2006.
Improving statistical MT by coupling reordering and decoding.
Machine Translation, 20(3):199–215.


George Foster and Roland Kuhn.
2007.
Mixture-model adaptation for SMT.
In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic.


Jianfeng Gao and Xiaodong He.
2013.
Training mrf-based phrase translation models using gradient ascent.
In *Proceedings of NAACL-HLT*, pages 450–459.


Mark Hopkins and Jonathan May.
2011.
Tuning as ranking.
In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK., July.


Hai-Son Le, Ilya Oparin, Alexandre Allauzen, Jean-Luc Gauvain, and François Yvon.
2011.
Structured output layer neural network language model.
In *Proceedings of ICASSP*, pages 5524–5527.


Hai-Son Le, Alexandre Allauzen, and François Yvon.
2012.
Continuous space translation models with neural networks.

In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 39–48, Montréal, Canada.

Rico Sennrich.
2012.
Perplexity minimization for translation model domain adaptation in statistical machine translation.
In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549.

Patrick Simianer, Stefan Riezler, and Chris Dyer.
2012.
Joint feature selection in distributed stochastic learning for large-scale discriminative training in SMT.
In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11–21. Association for Computational Linguistics.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki.
2007.
Online large-margin training for statistical machine translation.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Citeseer.