Latent Domain Translation Models in Mix-of-Domains Haystack

Hoang Cuong and Khalil Sima'an

ILLC, University of Amsterdam

UNIVERSITEIT VAN AMSTERDAM

COLING 2014, Dublin, Ireland.

SMT with Mix-of-Domains Haystack

We have **Big DATA** to train SMT systems.

- Thanks to Europarl, UN, Common Crawl, ...

Wait ...

- Data come from **very different domains**.
- **How does this affect SMT system performance**?

Bigger data $\neq$ better system performance

Exemplifying the Mix-of-Domains Haystack effect

Experiment:

- **Train** SMT systems on a haystack.
  Incl. EuroParl, Common Crawl, UN, News Commentary, Software, Hardware, etc.
- **Test** on a specific test set related to **Electronics**.

Haystack = 1M training sentence pairs

Input: *Se puede crear un archivo autodescodificable cuando el archivo codificado se abre con la contraseña maestra.*

Reference: *A self-decrypting file can be created when the encrypted* **file** *is* **opened with** *the* **master password**.

Output: *To create an file autodescodificable when the* **file** *codified* **commenced with** *the* **password teacher**.

Why? Haystack = relevant + irrelevant translations!

maestra → **master** (**computer**);
maestra → **teacher** (**education**); maestra → **dean** (**education**);
maestra → **crack** (**other**), maestra → ...

Haystack = 2M training sentence pairs

Input: *El reproductor puede reproducir señales de audio grabadas en mix-mode cd, cd-g, cd-extra y cd text.*

Reference: *The player* **can play back audio signals** *recorded in mix-mode cd, cd-g, cd-extra and cd text.*

Output: *The player* **can reproduce signs of audio** *recorded in mix-mode cd, cd-g, cd-extra and cd text.*

Why? Haystack = relevant + irrelevant translations!

reproducir → **reproduce** (**science**);
reproducir → **play back** (**electronics**);
reproducir → **render** (**software**);
reproducir → ...

Haystack = 4M training sentence pairs

Input: *Repite todas las pistas (únicamente cds de vídeo sin pbc)*

Reference: *Repeat* **all tracks** *(non-pbc video cds only)*

Output: *Repeated* **all avenues** *(only cds video without pbc)*

Why? Haystack = relevant + irrelevant translations!

pistas → **tracks** (**electronics**);
pistas → **runway** (**news**);
pistas → **avenues** (**wrong alignment**);
pistas → **arena** (**electronics**);
pistas → ...

The Haystack Challenge

Problem: Including **irrelevant data** might be **more harmful than beneficial**!

Challenge: **How to select the relevant sentences**?

Data Selection

Problem statement

Given an in-domain task:

- Input: A **haystack** (C_{mix}), an **in-domain** data set (C_{in}).
  - Role of C_{in}: **exemplifies** the task.
- Output: A subset of **relevant** data for the task.

Motivation

Data	**Phrases**	**BLEU**
4.6**M pairs**	236.74M	36.8
Select 300**K pairs**	22.47M	37.7/**+0.9**

Previous Work in the Literature

Perplexity-based selection methods based on **monolingual** language models between C_{in} and C_{mix}:

- Typical models: [Moore and Lewis(2010)], [Axelrod et al.(2011)Axelrod, He, and Gao], etc,.
- Idea: rank using monolingual cross-entropy difference over each side of corpus, or sum of them.

- Why only monolingual LMs?
  - What about translation differences between domains?
- Is there a more interesting way to model the relevance/irrelevance of sentences?

Examples

Each word or phrase might translate in different ways w.r.t different domains.

- maestra $\rightarrow$ **master**; (probably **software**, or **hardware**)
- maestra $\rightarrow$ **teacher**/**dean**; (probably **education**)
- maestra $\rightarrow$ **crack**, (probably **others**)...

Inviting Sentence Pairs: Relevance

Probabilistically recruit sentence pairs $\langle \mathbf{e}, \mathbf{f} \rangle$ that:

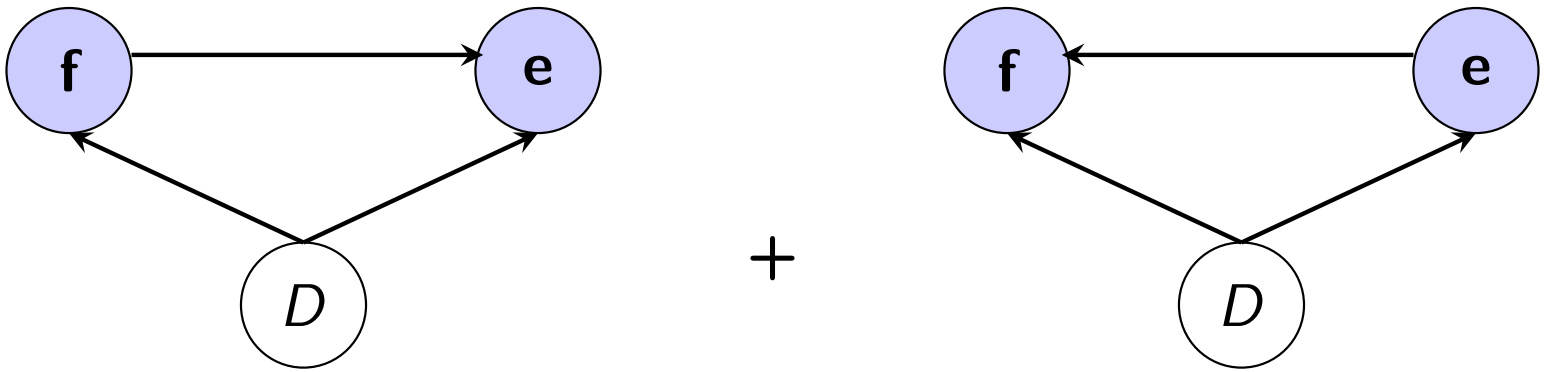
- to be in-domain, **and**,
- to be out-domain.

Only a little Math over the Generative Process ...

- A **latent domain variable** D is introduced to represent domains
  - D_1 means in-domain,
  - D_0 means out-domain.
- A **generative process** of sentence pairs $\langle \mathbf{e}, \mathbf{f} \rangle$ over D:

$$P(\mathbf{e}, \mathbf{f}|D).$$

Only a little Math over the Generative Process ...



$$P(\mathbf{e}, \mathbf{f} \mid D) = \frac{1}{2} \times \{P_{lm}(\mathbf{e} \mid D) P_t(\mathbf{f} \mid \mathbf{e}, D) \; + \; P_{lm}(\mathbf{f} \mid D) P_t(\mathbf{e} \mid \mathbf{f}, D)\}$$

- $P_{lm}(\mathbf{e} \mid D)$, $P_{lm}(\mathbf{f} \mid D)$: domain-dependent **language** models.
- $P_t(\mathbf{e} \mid \mathbf{f}, D)$, $P_t(\mathbf{f} \mid \mathbf{e}, D)$: domain-dependent **translation** models.
  - For simplicity, the model is induced at **word-alignment level**, and inspired by IBM Model 1.

In the end...

From $P(\mathbf{e}, \mathbf{f}|D)$, we can induce a "relevance" model $P(D|\mathbf{e}, \mathbf{f})$:

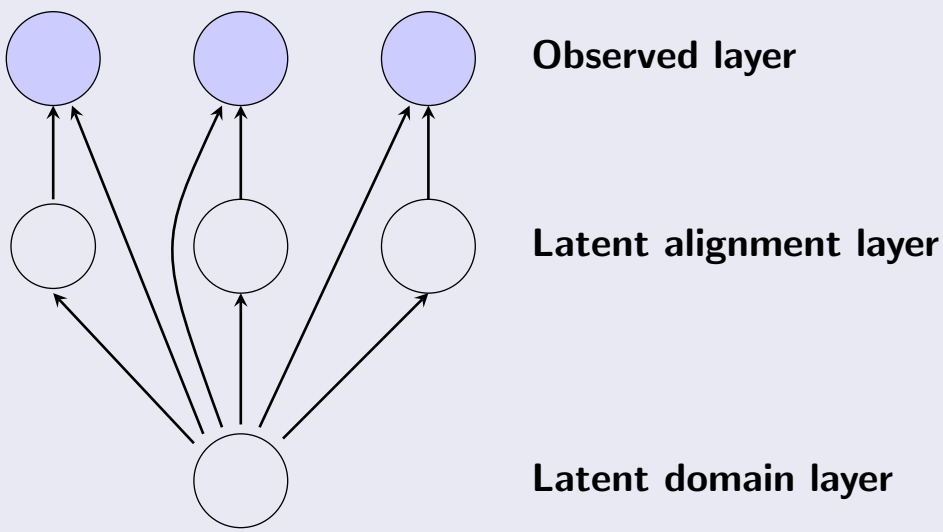
$$P(D|\mathbf{e}, \mathbf{f}) \propto P(\mathbf{f}, \mathbf{e}|D)P(D)$$

- With $P(D|\mathbf{e}, \mathbf{f})$:
  - Rank $P(D_1|\mathbf{e}, \mathbf{f})$, select most relevant (**This work**)
  - Train an SMT system in biased-style (instance weighting) (**Our work at EMNLP 2014**)

Training

Two kinds of latent variables: the **alignment** and **domain**.

- For simplicity, we fixed language models once trained.
- Parameters: the four **domain-conditioned word translation tables**.



Training

Train both latent variables (alignment and domain) *simultaneously*, using **Expectation Maximization**.

1. Find pseudo out-domain sample C_{out} (**"burn-in" process**).
   - Use TM from C_{in} only and uniform otherwise: No LMs!
   - Re-estimate: Single EM iteration
   - Select lowest scoring $P(D = 1 \mid \mathbf{e}, \mathbf{f})$ (same size as C_{in}).
2. Start actual EM training initialized as follows:
   - Use pseudo C_{out} and C_{in} to train LMs: Not C_{mix}!
   - Initialize TMs from C_{out} and C_{in}.

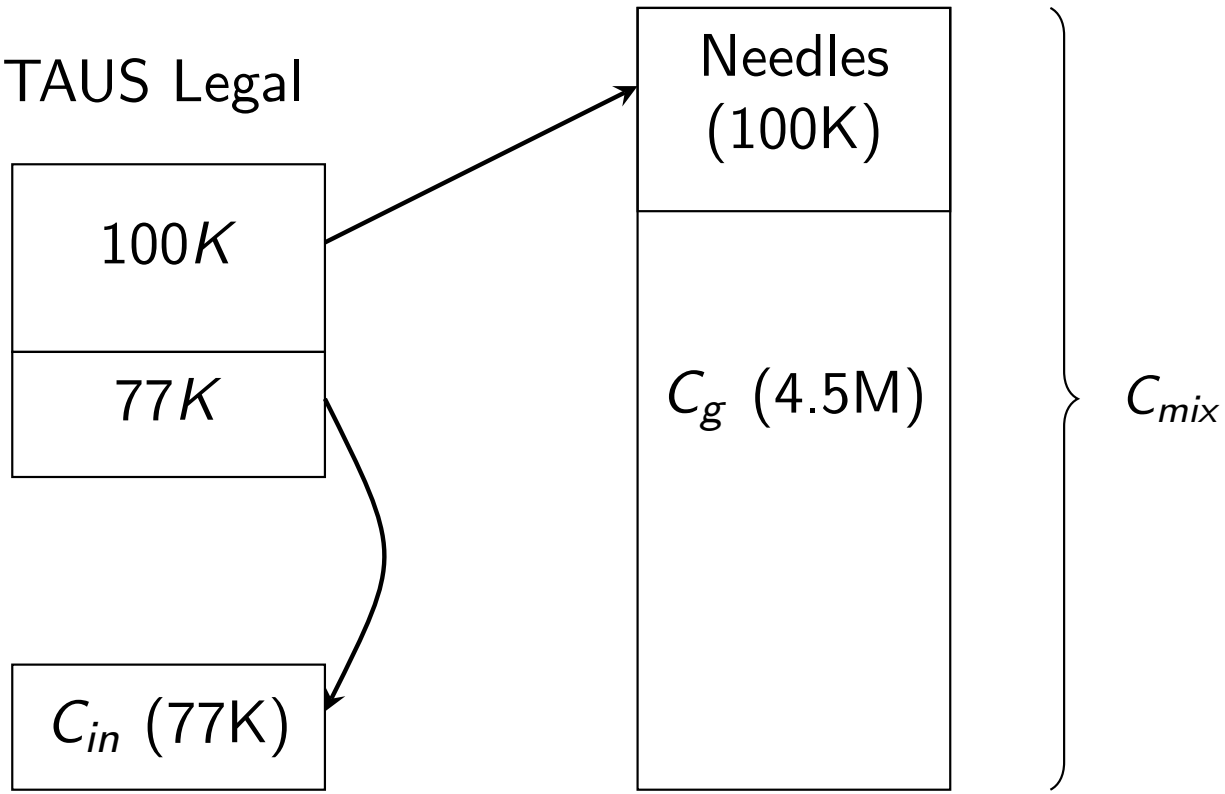
More rigorous Maths and details can be found in the paper.

Data Preparation

- General-domain data set C_g: $4.5M$ sentences pairs, collected from multiple resources
  - EuroParl, Common Crawl, UN, News Commentary, Software, Hardware, and Pharmacy.
  - A large and **rather varied set of domains haystack**!!!
- An in-domain data set:
  - $177K$ in-domain sentence pairs.
  - Domain: TAUS Legal.[1]

[1] Thanks to Translation Automation Society (TAUS.com)

Data Preparation



Evaluation

- Implementation: The code can be downloaded at: `https://sites.google.com/site/hoangcuong2011/`
- Evaluation Strategy:
  - *Intrinsic Evaluation*: (**new proposed strategy!**)
    - How many needles are among the tops?
  - *Extrinsic Evaluation*:
    - Translation yielded by systems trained on selected subsets?

Results

Model	**Needles**	**BLEU**
CE Difference (source side)	1122	28.2
CE Difference (target side)	1093	26.4
Bilingual CE Difference	1169	27.8
Invitation	53892	37.7
C_{mix}	-	36.8

What can we learn from the results?

- **Perplexity-based** methods fail in a noisy haystack.
- **Invitation model** works superior in these cases.

Results

Model	Needles	BLEU
Only LMs	34156	35.8
Only TMs	51991	37.4
Full Model	53892	37.7

Which is the key in the performance of Invitation Model?

- Translation models are crucial for performance.
- Language models make a small, yet noteworthy contribution.

Results

Model	Needles	BLEU
Our Model with only LM	34156	35.8
Best Baseline	1122	28.2

Which is better way to derive LMs?

- in- and out-domain data? (**novel**) v.s.
- in- and mix-domain data? (**common approach**)

Deriving separately in- and out-domain LMs produces much better in this case!

This also reasons why previous works failed

- We introduce a new take on the selection problem.
  - A generative process of data over hidden variables has been proposed.
  - A reminiscent EM algorithm has been designed.
  - Superior performance compared to previous models.
  - More details about the model, and experimental results can be found on our paper.
- The problem is expanded to not limited at data selection only, e.g., see our work at EMNLP 2014 [Cuong and Sima'an(2014)].

- Appeared in a technical report by:

  Amir Kamran, Bart Mellebeek and Khalil Sima'an

- Note: a much "easier" test case!!!

Cut-offs	**Model**	**Precision**	**BLEU**
$50K$	Bilingual CE Difference	0.445	34.90
	Invitation	0.836	45.31
$100K$	Bilingual CE Difference	0.453	40.45
	Invitation	0.748	45.84
$300K$	Bilingual CE Difference	0.512	46.40
	Invitation	0.557	46.62
C_{mix}	–	–	46.31

Table: Further results: C_{in} = TAUS Software 100K, C_{mix} = $1.9M$, consisting of Taus Legal 400K + Common Crawl 400K + Taus Mix 400 K + Taus Software 300K.

- Appeared in a technical report by:

  Amir Kamran, Bart Mellebeek and Khalil Sima'an

- Note: a much "easier" test case!!!

Cut-offs	**Model**	**Needles**	**BLEU**
$50K$	Bilingual CE Difference	0.621	37.05
	Invitation	0.910	40.41
$100K$	Bilingual CE Difference	0.632	40.16
	Invitation	0.881	41.94
$300K$	Bilingual CE Difference	0.617	42.81
	Invitation	0.729	43.23
C_{mix}	–	–	43.30

Table: Further results: C_{in} = TAUS Legal 100K, C_{mix} = $1.9M$, consisting of TAUS Software 400K + Common Crawl 400K + TAUS Mix 800K + TAUS Legal 300K.

Bibliography I

Amittai Axelrod, Xiaodong He, and Jianfeng Gao.

Domain adaptation via pseudo in-domain data selection.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
ISBN 978-1-937284-11-4.
URL http://dl.acm.org/citation.cfm?id=2145432.2145474.

Hoang Cuong and Khalil Sima'an.

Latent domain phrase-based models for adaptation.
In *EMNLP 2014*, 2014.
To appear.

Robert C. Moore and William Lewis.

Intelligent selection of language model training data.
In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
URL http://dl.acm.org/citation.cfm?id=1858842.1858883.