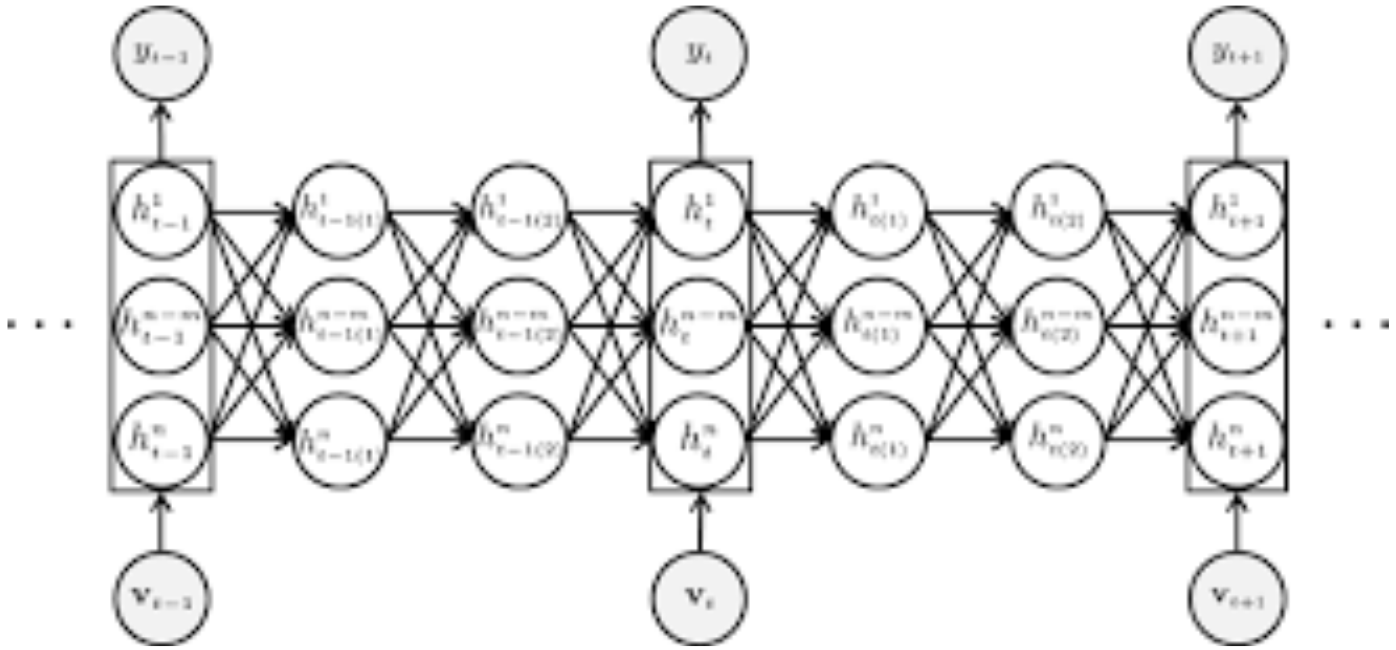
USERS AND DATA: THE TWO NEGLECTED CHILDREN OF BILINGUAL NATURAL LANGUAGE PROCESSING RESEARCH

Philippe Langlais, RALI, Université de Montréal

joint work with Julien Bourdaillet, Alexandre Bérard, Fabrizio Gotti, Francis Grégoire, Stéphane Huet, Laurent Jakubina, Fethi Lamraoui, Guy Lapalme, Alexandre Patry, Lise Rebout

NLP Research Today ...



Models, models, models, models
(preferably deep ones)

But User Matters too

- Not much reflected in current evaluation protocols
  - e.g. how useful is a model that can identify the translation of **frequent words** at an accuracy of 40% ?
- Smart evaluations do exist though
  - (Sharoff et al., 2006) for translating difficult MWEs

TRANSSEARCH 3 BETA TERMINOTIX rali

UTILISATEUR : lapalme | REQUÊTES | MON COMPTE | PRÉFÉRENCES | AIDE | QUITTER

Signet / Favori personnalisé : TransSearch (qu'est-ce que c'est ?) Requête bilingue

Collection de documents : Les Hansards canadiens

Expression : take+ .. ride Chercher

92 traductions de *take+ .. ride* dans 106 occurrences

dindons de la farce	4
monté un bateau	3
faire avoir	3
se fasse rouler	2
fait berner	2
se fait jouer	2
moqués de	2
fait	2
les a	2
se sont fait avoir	2
le public pour attirer la	1
a fait une ballade	1
nous rouler dans ce projet nous tous	1
en train de monter un bateau à la population canadienne	1
tête des contribuables que se paie le	1
passer une petite vite	1
bourrer de l'autre côté de la chambre en	1
ont pris la voiture que pour faire une balade	1

dindons de la farce 4

Emissions continue to rise and taxpayers are being **taken along for the ride**.	Les émissions continuent d'augmenter et c'est le contribuable qui est **le dindon de la farce**.
They are left with nothing. Now they are here illegally with no documentation. Canadians are being **taken for a ride**.	Ces personnes se trouvent ici illégalement, elles n'ont aucun document et nous, les Canadiens, sommes les **dindons de la farce**.
This would affect close to 400,000 Canadians, 80,000 of them Quebecers, who have been the ones **taken for a ride**.	Il s'agit d'une mesure qui toucherait près de 400 000 Canadiens, dont 80 000 Québécois, qui ont été les **dindons de la farce**.
I think that this is a prime example of a tainted system in which people who cannot afford to invest in sectors eligible for tax credits are urged to do so through all kinds of scams and end up being **taken for a ride**.	Je pense que c'est un exemple patent d'un système vicié, où des gens qui n'ont pas les moyens d'investir dans des domaines où on peut obtenir des crédits d'impôt se voient, par toutes sortes de subterfuges, invités à le faire et, en bout de ligne, ils se trouvent à être **les dindons de la farce**.

TransSearch

- Users are tolerant to alignment problems
  - We implemented a detector of erroneous alignments but decided not to integrate it
  - We tested a number of alignment models but used one of the simplest (IBM model 2 with contiguity constraint)
- Users care more about translation diversity in the top candidate list
  - so clustering translations is of better use for a user
- See (Bourdaillet et al., 2010)

File Edit Go Options About

Product overview

The machine is controlled from a liquid crystal color touch screen where you can view your operation and application settings.

Printer settings can be pre-programmed for specific production job types and when such a job type is selected, the printer is set up automatically for the paper type and application.

Aperçu de la machine:

La machine est contrôlée à partir d'un écran tactile à cristaux liquides vous permettant de visualiser vos paramètres d'application et d'opération.

Les paramètres de l'imprimante peuvent être préprogrammés pour des types de type de travail et lorsqu'il est sélectionné

rs de la
rsqu'il est sélectionné,
rsque le le
rsque le,

TransType

- Keystroke saving is a poor measure of the usefulness of the tool
  - often, users do not watch the screen… (Langlais et al., 2002)
- Cognitive load for reading completions
  - predicting when a user will find a completion useful is a plus (Foster et al., 2002)
- Users do not like repetitive errors
  - adapting online is important (Nepveu et al., 2004)

See (Gonzales-Rubio et al., 2012) for advances in targeted mediated interactive MT

And Data Matters too

- Often the main issue
  - once you have data, use your ``best hammer''
  - e.g.
    - domain specific MT
    - organizing a dataflow
- **Know-how** in handling Users and Data is in need in practical settings
  - but not much rewarded academically

Of course …

- Better models are important !
  - e.g. NMT (Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2014)
    - simpler (single model)
    - better results
      - (Bentivogli et al., 2016; Isabelle et al.; 2017)
    - good properties for further improvements
      - multilingual
      - continuously learning
      - multi-tasking, etc.
- And have the potential to impact users

Plan

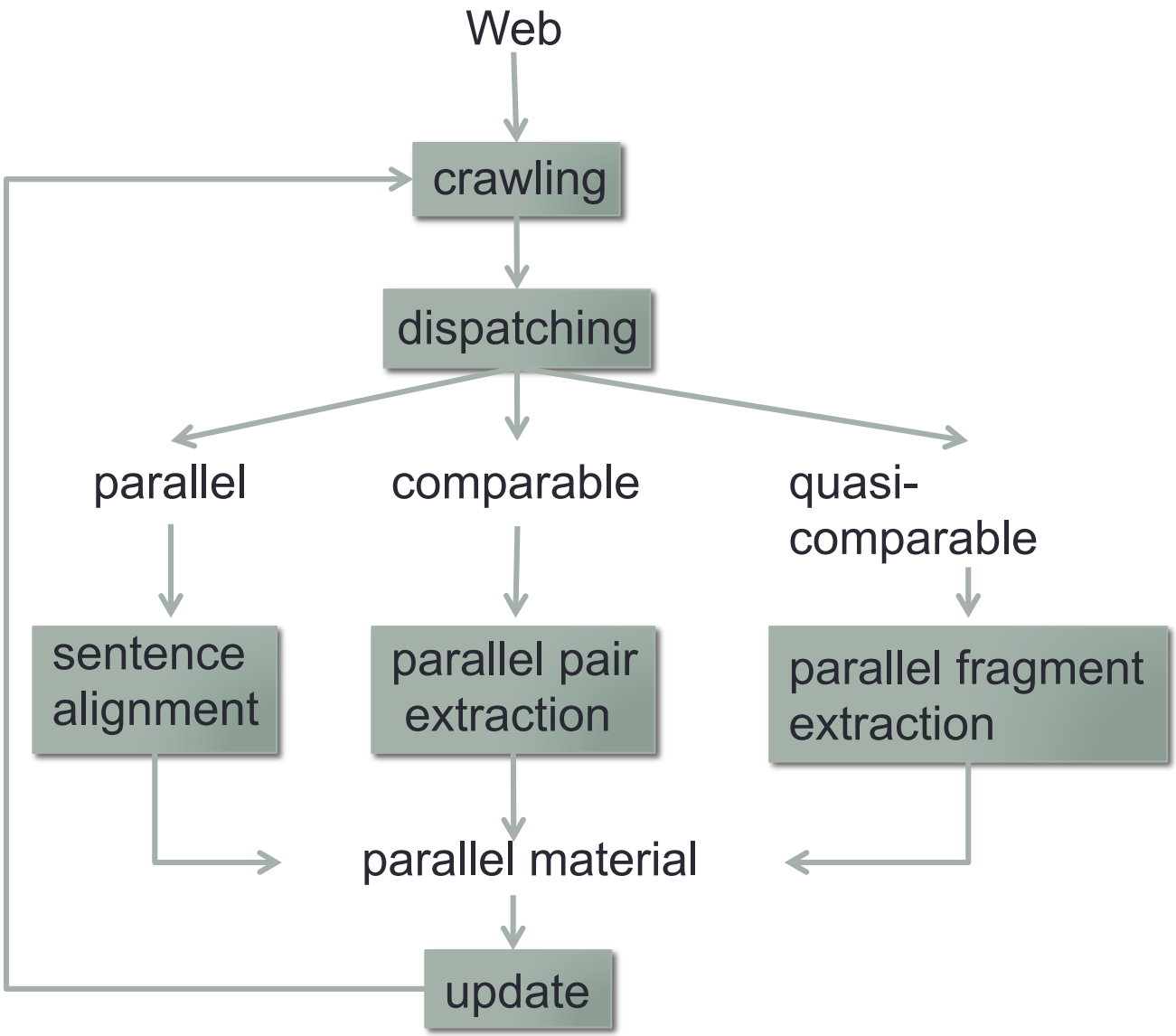
① Data Management (Acquisition and Organization)

- What has been done
- What (I feel) is still missing
  - **Domain specific** bilingual parallel and comparable corpora
  - Assessing the **quality/usefulness** of a pair of documents

② Parallel Material Extraction from a Comparable Corpus

- Document pairs
- Sentence pairs
- Word pairs

Data Management Overview



Input

- seed terms/urls
- documents of interest (mono or bilingual)

Objective

- e.g. improving a generic SMT engine

Data Management: What has been done

- Web2ParallelDoc/Sent
  - BITS (Ma & Liberman, 1999)
    - a sensible pipeline, leveraging lexicons, cognates and a list of webdomains
  - PTMiner (Chen & Nie, 2000), STRAND (Resnik & Smith, 2003)
    - rule-based URL matching, HTML-tag and lexicon-based pairing
  - (Nadeau & Foster, 2004; Fry 2005)
    - specific to newswire feeds
  - WebMining (Tomas et al., 2005)
    - pipeline which only requires a list of bilingual websites (dictionaries are trained online)
  - (Fukushima, et al., 2006)
    - leveraging a dictionary2graph algorithm, and paying attention to speed issues
  - BABYLON (Mohler and Mihalcea, 2008)
    - low-density language oriented, a pipeline which input is a source document
  - BITEXTOR (Espla-Gomis et Forcada, 2010)
    - pipeline of URL- and content-based matching

- See (Kulkarni, 2012) for a survey

Data Management: What has been done

- Notable Large-scale Efforts
  - (Callisson-Burch et al., 2009)
    - 10^9 word parallel corpus crawled from specific URLs, URL-based pairing, sentence alignment (Moore 2002) + cleaning: 105M sentence pairs
  - (Uszkoreit et al., 2010)
    - parallel document detection using MT and near-duplicate detection
    - 2.5B webpages crawled from the Web + 1.5M public-domain books
    - 24 hours on a 2000-node cluster
  - (Ture & Lin, 2012)
    - English-German parallel sentences extracted from Wikipedia tackling the cross-lingual pairwise similarity problem without heuristics
  - (Smith et al., 2013)
    - Open-source extension of STRAND, impressive deployment on the CommonCrawl corpus (32TB), using Amazon's elastic map-reduce

Data Management: Issues

- No **systematic comparison** of those systems
- Not much sensibility to **specific domains**

- Crawling
  - which input? manual set of webdomains / urls / terms ?
  - iterative or not ?
  - concerns: efficiency, coverage, resources expected

- **Dispatching**
  - grade a pair of documents (Li and Gaussier, 2010; Fung and Cheung, 2004; Babych and Hartley, 2014)
  - measure of the usefulness of a resource (Barker and Gaizauskas, 2012)

- Bootstrapping
  - model/lexicon update (Fung and Cheung, 2004)

Data Management: Issues

- Experiments in those directions would lead to :
  - a better know-how
    - valuable in practical (industry) settings
  - a shared repository of domain specific collections
    - as OPUS for parallel texts (Tiedmann, 2012)
  - (potentially) a reference bilingual collection where all parallel units are marked
    - would help measuring progress
    - ease reproducibility

About Domain Specific CC

- Domain specific crawling (monolingual)
  - WebBootCat (Baroni and Bernardini, 2004)
  - TerminoWeb (Barrière and Agbago, 2006)
  - Babouk (De Groc, 2011)
  - (Azoulay, 2017) considering only pdf files
- Specialized (bilingual) comparable corpora acquisition
  - See the ACCURAT project (Pinnis et al., 2012; Aker et al., 2012)
  - METRICC (Alonso et al., 2012)
  - Not just a matter of quantity (Morin et al., 2007)

Plan

① Data Management (Acquisition and Organization)

- What has been done
- What (I feel) is still missing
  - **Domain specific** bilingual parallel and comparable corpora
  - Assessing the **quality/usefulness** of a pair of documents

② Parallel Material Extraction from a Comparable Corpus

- Document pairs
- Sentence pairs
- Word pairs

Identifying parallel documents

- Has received some attention (see previous slides)
  - (Enright and Kondrak, 2010)
    - nb of shared hapax words
  - (Uszkoreit et al., 2010)
    - machine translation + monolingual duplicates detection
  - (Patry and Langlais, 2011)
    - light doc. representation + IR + classifier
- Often just one component of a pipeline which is not evaluated as such

Paradocs (Patry and Langlais, 2011)

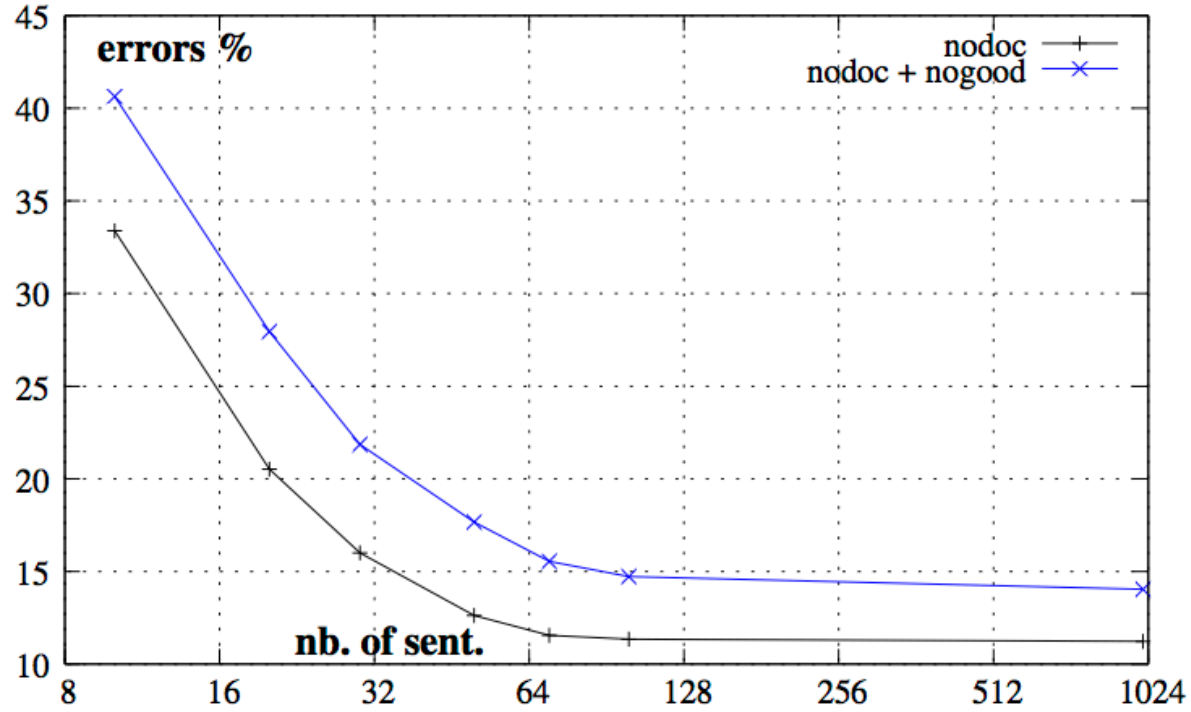
- **Light** way of representing a document:
  - sequence of numerical entities
  - sequence of hapax words (of at least 4 symbols)
  - sequence of punctuations . ! ? ( ) :

  as in (Nadeau and Foster 2004)

- Avoid Cartesian product thanks to IR
  - **no need for a bilingual lexicon**

- Train a classifier to recognize parallel documents
  - normalized edit-distance between each sequence representation
  - count of each entity (numerals, punctuations, hapaxes)

Paradocs: Avoiding the Cartesian product



Europarl setting
- 6000 bitexts per language pair
- Dutch2English

- nodoc: no document returned by Lucene
- nogood: no good document returned (n=20)

- ➢ >40% of failure for very short documents (<10 sentences)
- ➢ 15% for long documents (>64 sentences)

WMT 2016 Shared Task on Bilingual Document Alignment

- Goal: better understanding best practices
  - 11 teams, 21 systems submitted (+ 1 baseline)

- Easy entrance task:
  - train: 49 webdomains crawled (ex: virtualhospice.ca)
    - only HTML pages considered
    - text pre-extracted, duplicates removed, language pre-identified (en-fr), machine translation provided
  - test: 203 other webdomains
  - evaluation:
    - retrieving 2400 pairs of documents known parallel (within webdomains)
    - strict rule: a target document should be proposed at most once

RALI's participation (Jakubina and Langlais, 2016)

- motivation: simplicity !
  - no use of MT, eventually even no lexicon-based translation
- an IR approach (Lucene-based) comparing/combining
  - IR (monolingual)
  - CLIR (based on a bilingual lexicon for « translating »)
  - a simple but efficient URL-based IR (tokenizing URLs)
- `badLuc` ended up with a recall of 79.3%
  - the best system recorded 95%
  - the organizers' baseline 59.8%

Results on train

Strategy	@1
text monolingual	
default	6.4
default+tok	35.4
best (w/o length)	64.9
best (w length)	76.6
text bilingual	
best (w length)	83.3
url	
baseline WMT16	67.9
best	80.1
badLuc	
w/o post-treat	88.6
w post-treat	92.1

- Playing with metaparameters helps a lot
- Applying a length filter also helps a lot

- Involving translation is a must
  - even with a simple lexicon (which covers ~half of the words in the collection)

- Our URL variant is performing impressively well
  - outperforms the baseline
  - useful on short documents

- Combining both indexes (text and urls) helps
- Post-filtering is a plus

Plan

① Data Management (Acquisition and Organization)

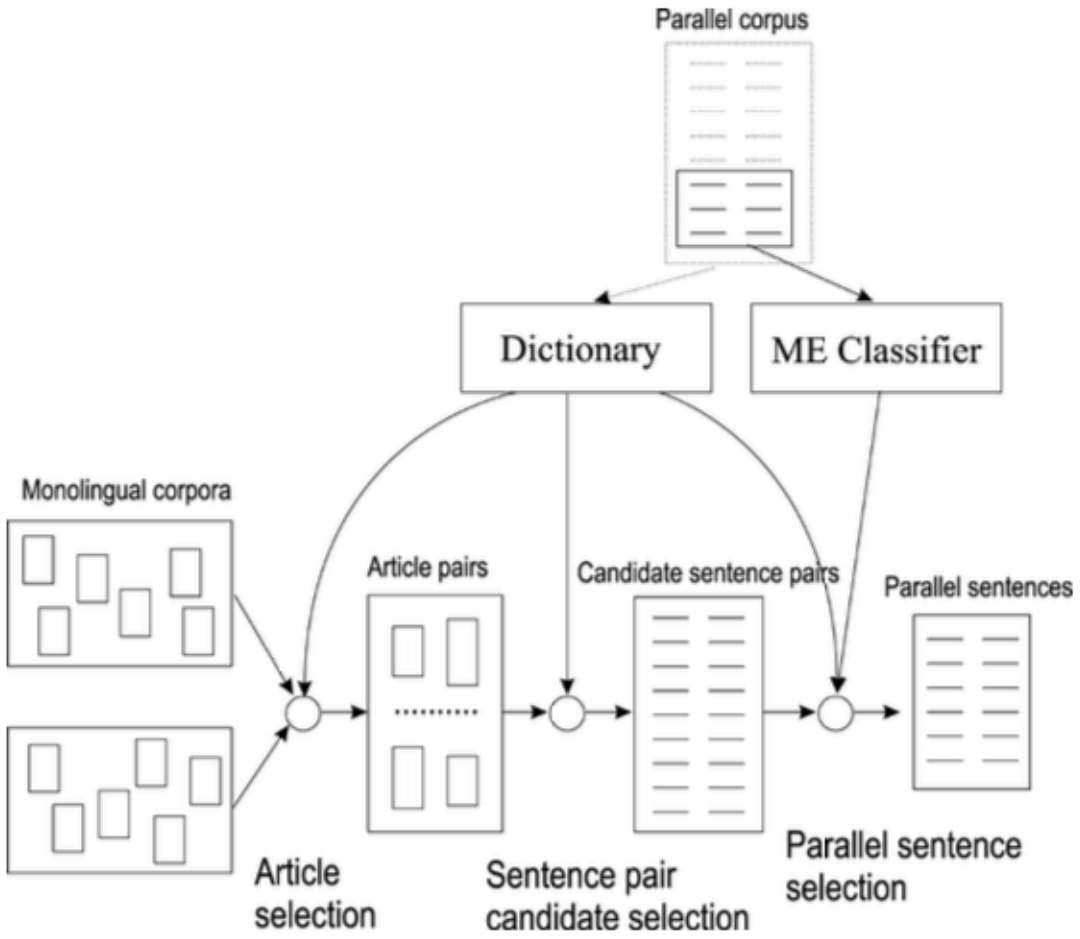
- What has been done
- What (I feel) is still missing
  - **Domain specific** bilingual parallel and comparable corpora
  - Assessing the **quality/usefulness** of a pair of documents

② Parallel Material Extraction from a Comparable Corpus

- Document pairs
- Sentence pairs
- Word pairs

MUNT (Bérard, 2014)

A reimplementation of (Munteanu and Marcu, 2005) with features borrowed from (Smith et al., 2010)



MUNT (Bérard, 2014)

- classifier (logistic regression)
  - 31 features
    - length based, alignment-based features, fertility, etc.
- pre-filter (selecting interesting sentence pairs)
  - ratio of sentence length no greater than 2
  - at least 50% of tokens with alignment in the other side
  - removes ~98% of the Cartesian product !!

MUNT on Wikipedia

- Done before (Smith et al., 2010)
- Wikipedia dump (2014)

	#articles	#paired	#sent.	#tokens
en	4.5M	919k	29.3M	630.8M
fr	1.5M	919k	16.8M	354.5M

- Configuration (default)
  - lexicon trained with GIZA++ on 100k sentence pairs of Europarl
  - classifier trained on 800 positive and 800 negative examples of news data, threeshold: 0.8

MUNT on Wikipedia

Manually evaluated 500 (random) sentence pairs

- parallel

- quasi-parallel (at least partial)

At the cellular level, the nervous system is defined by the presence of a special type of cell, called the neuron, also known as a ”nerve cell” .

À l’échelle cellulaire, le système nerveux est défini par la présence de cellules hautement spécialisées appelées neurones, qui ont la capacité, très particulière, de véhiculer un signal électrochimique .

- not parallel

MUNT on Wikipedia

grade	%
parallel	48
quasi-parallel	16
not parallel	36

- 15 hours (on a cluster of 8 nodes)
- MUNT detected 2.61M sentence pairs
- 2.26M once duplicates removed
- 1.92M after removing sentences shorter than 4 words
- 64% of sentence pairs are (quasi-)parallel

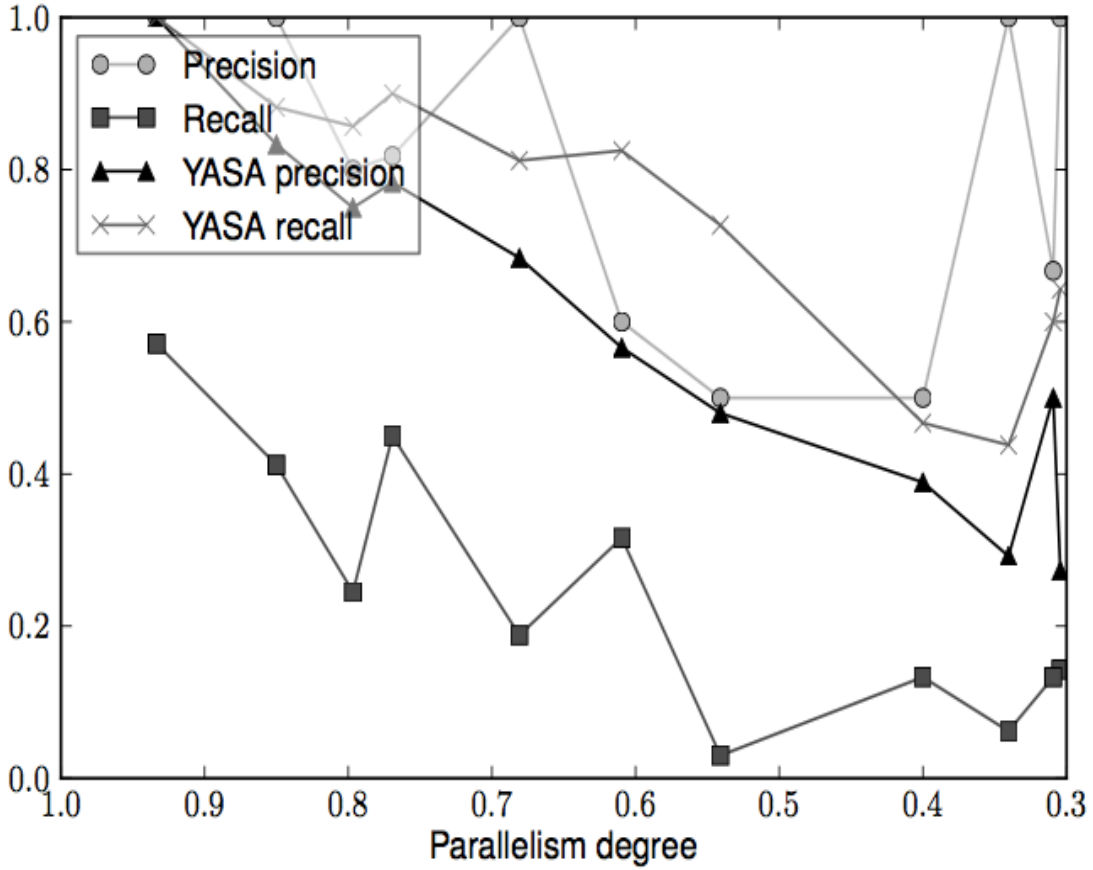
We tried something odd:

- We applied Yasa (Lamraoui and Langlais, 2013) on Wikipedia article pairs (pretending they were parallel)
- Asked MUNT to classify the sentence pairs identified

grade	%
parallel	71
quasi-parallel	15
not parallel	14

- 11M sentence pairs identified by Yasa
- 1.6M kept parallel by MUNT
- 1.3M once duplicate removed
- **86% of sentence pairs are (quasi-)parallel**
- much faster !

Munt on Wikipedia



- 50 comparable Wiki article pairs manually aligned at the sentence level (Rebout and Langlais, 2014)
- measured the performance of YASA and MUNT on those articles

- MUNT has a better precision, but a lower recall

$$\frac{2 * n_{para}}{n_{fr} + n_{en}}$$

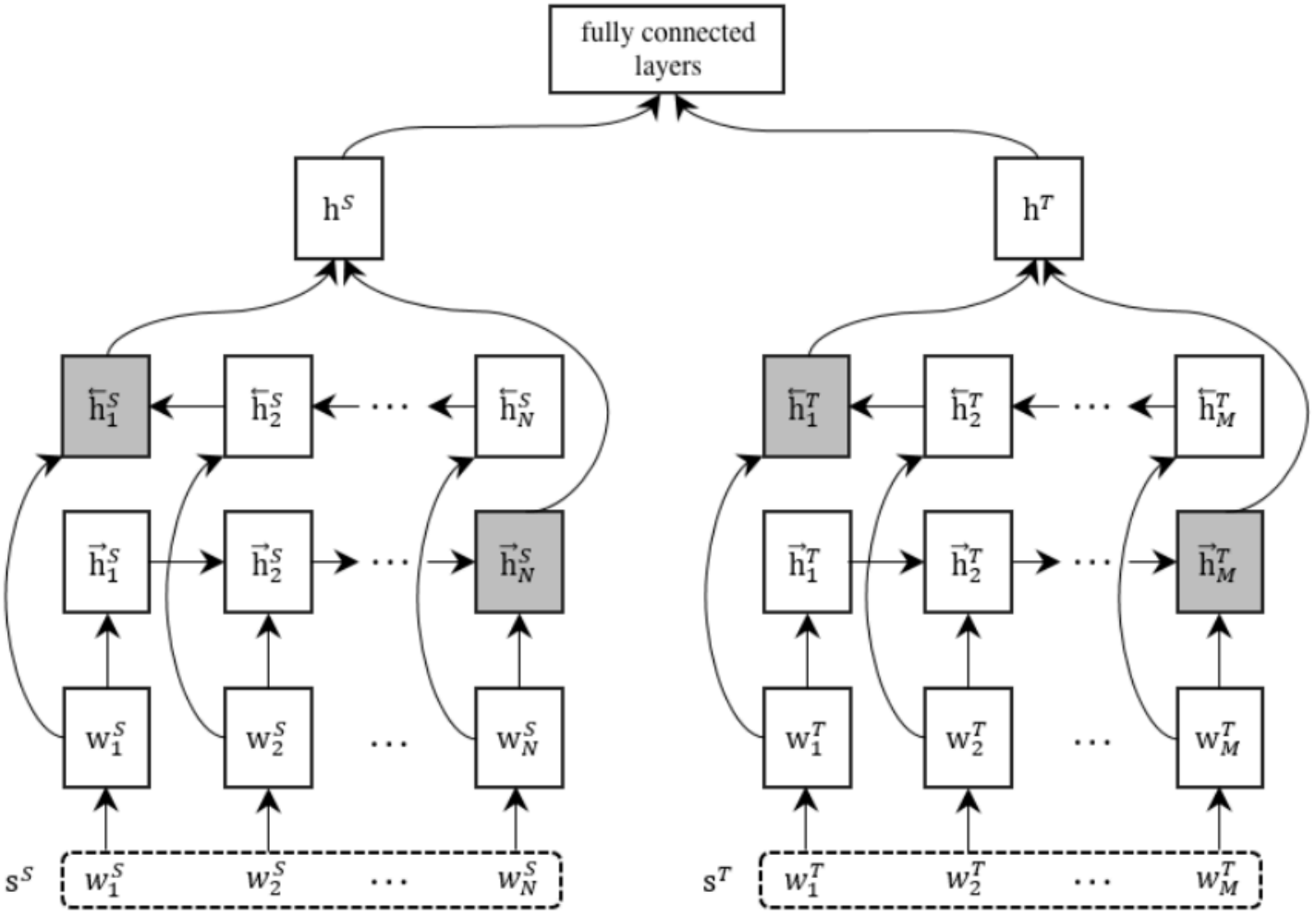
n_{para} # parallel sentences
n_{fr} (n_{en}) # sentences in fr (en) doc.

BUCC 2017 Shared Task (Zweigenbaum et al., 2016)

- Detecting parallel sentences in a large text collection
  - 2 sets of monolingual Wikipedia sentences (2014 dumps):
    - 1.4M French sentences
    - 1.9M English sentences
    - + 17k **parallel sentences** from News Commentary (v9)
  - Evaluation: precision, recall, F1

- pros
  - no metadata (text-based)
  - Cartesian product is large (with few positive examples)
  - Smartness in inserting parallel sentences (to avoid simple solutions)
- cons
  - Artificial task
  - True parallel sentences in Wikipedia EN-FR are not known

RALI's participation



Will be presented this afternoon (Grégoire and Langlais, 2017)

RALI's participation

- Training: Europarl v7 French-English
  - first 500K sentence pairs
  - negative sampling: random selection of sentence pairs
- Test: newstest2012 (out-domain)
  - 1000 first sentence pairs + noise
- Pre-processing
  - maximum sentence length: 80
  - tokenization with Moses' toolkit, lowercased
  - mapping digits to 0 (e.g. 1982 -> 0000)

RALI's participation

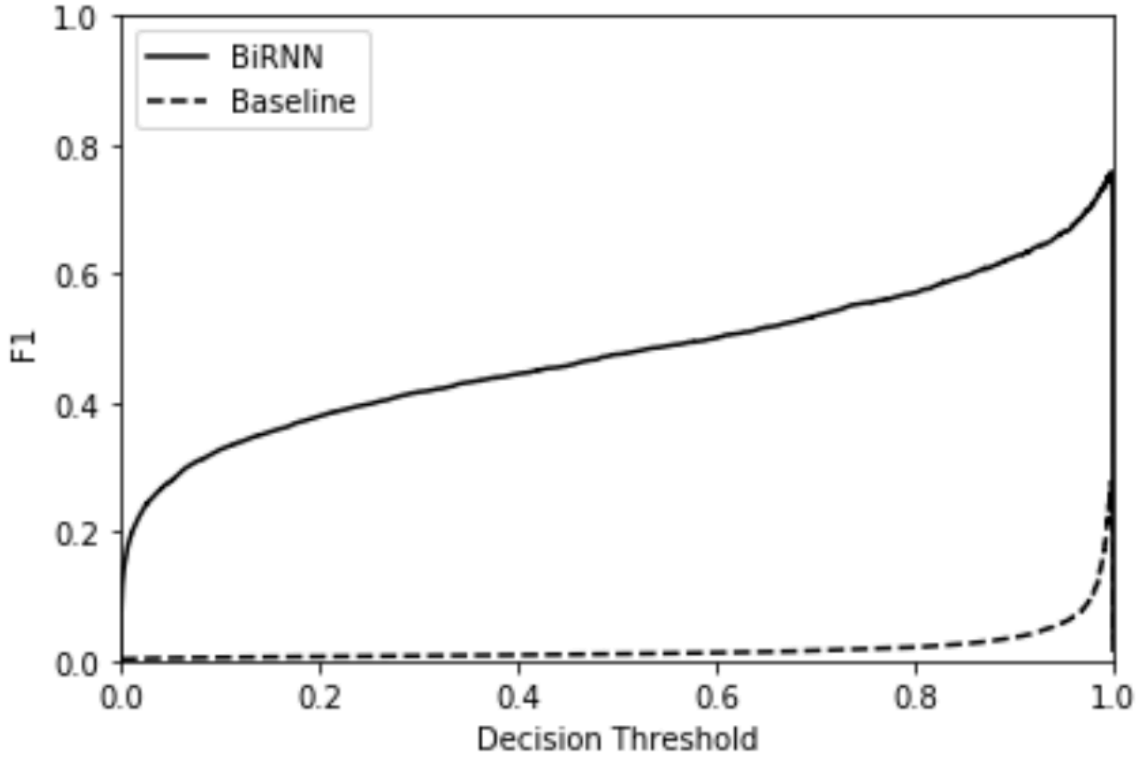
- Embedding-based filter for avoiding the Cartesian product
  - word-based embeddings computed with BILBOWA (Gouws, 2015)
  - sentence representation = average word embeddings
  - 40 first target sentences for each source one
- Not paying attention to ``details'' (only model matters, right?)
  - digit preprocessing
  - random negative sampling at training does not match testing condition

model	precision	recall	F1	
	12.1	70.9	20.7	*official*

- The next slides summarize what we have learnt **after** our participation to BUCC 2017 (new material)

Influence of the decision threeshold

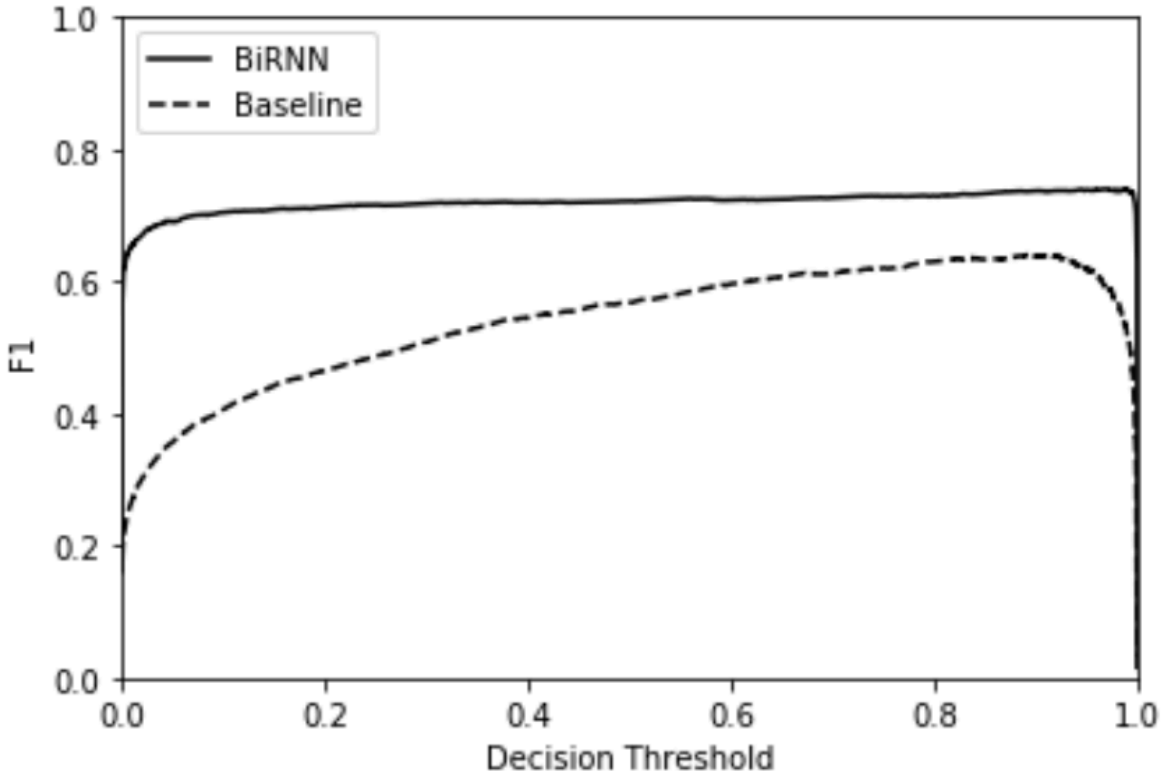
- Cartesian product
  - 1M ex., 1k positive
- BIRNN trained with 7 negative ex.
- MUNT trained with a balanced corpus



	Precision	Recall	F1	ρ
BiRNN	83.0	69.6	75.7	0.99
MUNT	31.8	24.1	27.7	0.99

Influence of the decision threeshold

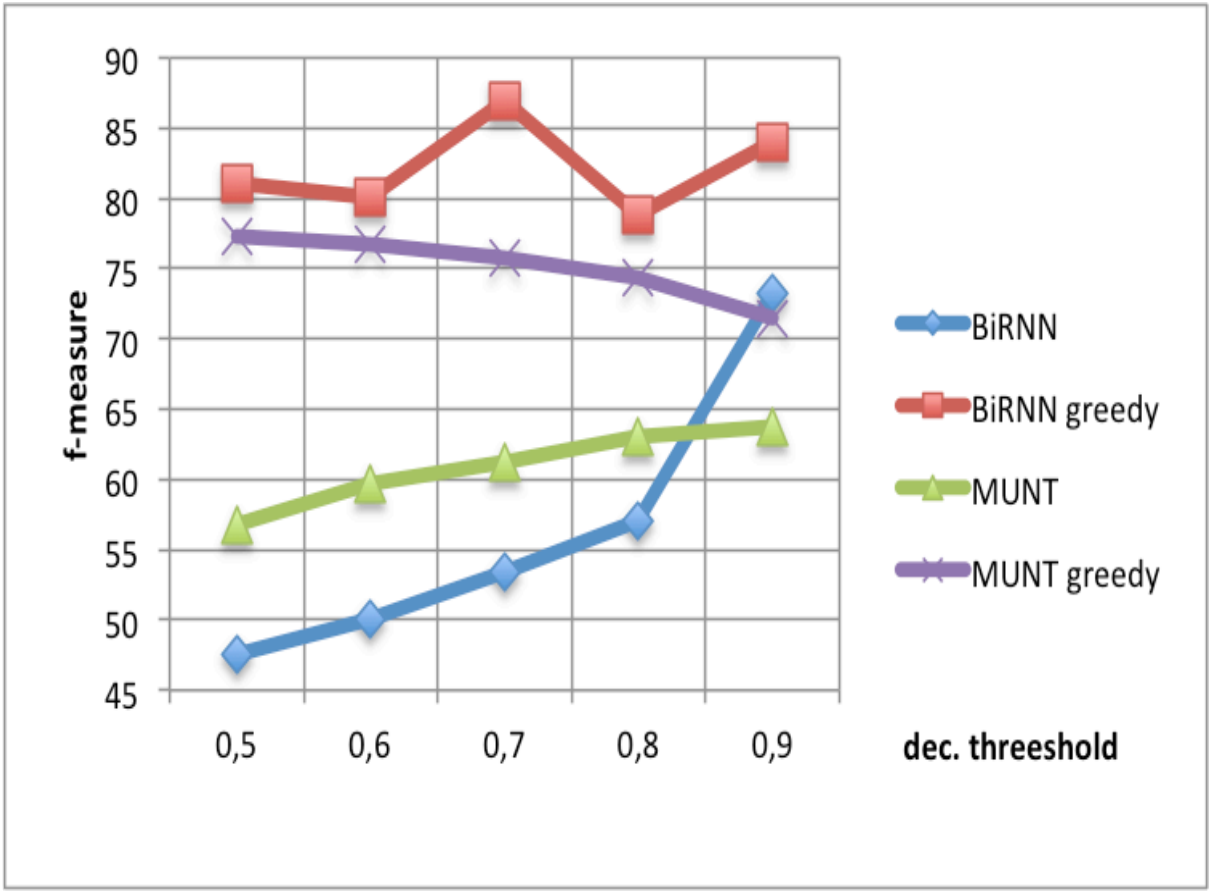
- Pre-filtering
  - 8053 ex., 1k positive
- BIRNN trained with 7 negative ex.
- MUNT trained with a balanced corpus



	Precision	Recall	F1	ρ
BiRNN	91.0	62.4	74.0	0.97
MUNT	73.3	57.0	64.1	0.91

Influence of Post-filtering

- Each decision taken independently
  - A src sent. may be associated to several target ones, and vice versa
    - ✧ post-filtering (greedy algorithm, Hungarian algo. too slow)



huge boost in precision at a small recall loss for both approaches

Plan

① Data Management (Acquisition and Organization)

- What has been done
- What (I feel) is still missing
  - **Domain specific** bilingual parallel and comparable corpora
  - Assessing the **quality/usefulness** of a pair of documents

② Parallel Material Extraction from a Comparable Corpus

- Document pairs
- Sentence pairs
- Word pairs

Bilingual Lexicon Induction

- Received a lot of attention
  - Pioneering works: (Rapp, 1995; Fung, 1995)
  - See (Sharoff et al., 2013)
- Revisited as a way to measure the quality of word embeddings
  - Seminal work of (Mikolov et al., 2013)
- Comprehensive comparisons
  - (Levy et al., 2014, 2017; Upadhyay et al., 2016)

Bilingual Lexicon Induction
(Jakubina & Langlais; 2016)

- We thoroughly revisited those approaches:
  - (Rapp, 1995)
  - (Mikolov et al. 2013)
    - training the projection matrix with the toolkit of (Dinu and Baroni, 2015)
  - (Faruqui and Dyer, 2014)
  - and a few others, but without success

- investigating their meta-parameters
- paying attention to the frequency of the terms
  - after (Pekar et al, 2006)
- showing their complementarity

Experiments

- Wikipedia (dumps of June 2013)
  - EN: 7.3M token forms (1.2G tokens)
  - FR: 3.6M token forms (330M tokens)
- Test sets
  - Wiki≤25 English words occurring at most 25 times in Wiki-EN
    - 6.8M such tokens (92%)
  - Wiki>25 English words seen more than 25 times in Wiki-EN
  - Euro-5-6k top frequent 5000 to 6000 words of WMT2011

	Frequency			
	min	max	avg	cov (%)
Wiki≤25	1	25	10	100.0
Wiki>25	27	19.4k	2.8k	100.0
Euro5-6k	1	2.6M	33.6k	87.3

Metaparameters explored

- Rapp

No adhoc filtering as usually done, so very time and memory challenging for the Rapp approach

(Prochasson and Fung, 2011)	this work
- ~20k document pairs	~700k ones
- target voc. 128k words (nouns)	3M words

We did apply a few filters:
- context vectors: 1000 top-ranked words
- 50k first occurrences of a source term

Best variant (per test set)

	@1	@5	@20
Wiki>25		*(ed@1 19.3)*	
Rapp	**20.0**	**33.0**	**43.0**
Miko	17.0	32.6	41.6
Faru	13.3	26.0	33.3
Wiki≤25		*(ed@1 17.6)*	
Rapp	**2.6**	4.3	7.3
Miko	1.6	**4.6**	**10.6**
Faru	1.6	2.6	5.0
Euro5-6k		*(ed@1 8.0)*	
Rapp	16.6	31.8	41.2
Miko	**42.0**	**59.0**	**67.8**
Faru	30.6	47.7	59.8

Reranking Candidate Translations (Jakubina and Langlais, 2016)

- Reranking shown useful in a number of settings
  (Delpech et al., 2012; Harastani et al., 2013; Kontonatsios et al., 2014)

- Trained a reranker (random forest) with RankLib
  - 700 terms for training, remaining 300 terms for testing
  - 3-fold cross-validation

- Light features for each pair (s,t):
  - frequency-based features
    - freq of s,t and their difference
  - string-based features
    - length of s and t, their difference, their ratio, edit distance(s,t)
  - rank-based features
    - score and rank of t in the native list
    - number of lists in which t appears (when several n-best lists are considered)

Reranking Individual n-best Lists

	Individual			1-Reranked		
	@1	@5	@20	@1	@5	@20
Wiki>25						
Rapp	20.0	33.0	43.0	36.3	48.8	53.8
Miko	17.0	32.6	41.6	**38.1**	**49.0**	**54.3**
Faru	13.3	26.0	33.3	34.3	44.0	47.9
Wiki≤25						
Rapp	2.6	4.3	7.3	8.6	9.4	10.2
Miko	1.6	4.6	10.6	**16.6**	**19.0**	**20.1**
Faru	1.6	2.6	5.0	7.9	8.7	8.9
Euro5-6k						
Rapp	16.6	31.8	41.2	34.6	48.6	51.9
Miko	42.0	59.0	67.8	**47.0**	**68.1**	**73.0**
Faru	30.6	47.7	59.8	41.2	58.0	66.0

Reranking several n-best Lists

	1-Reranked				n-Reranked		
	@1	@5	@20		@1	@5	@20
Wiki>25							
Rapp	36.3	48.8	53.8				
Miko	**38.1**	**49.0**	**54.3**	R+M	43.3	58.4	62.4
Faru	34.3	44.0	47.9	R+M+F	**45.6**	**59.6**	**64.0**
Wiki≤25							
Rapp	8.6	9.4	10.2				
Miko	**16.6**	**19.0**	**20.1**	R+M	18.9	22.0	23.6
Faru	7.9	8.7	8.9	R+M+F	**21.3**	**24.4**	**25.7**
Euro5-6k							
Rapp	34.6	48.6	51.9				
Miko	**47.0**	**68.1**	**73.0**	R+M	**49.5**	**68.7**	76.1
Faru	41.2	58.0	66.0	R+M+F	47.6	68.5	**76.2**

One-Slide Wrap up

- Hyp: Model-centric research somehow hides the value of:
  - know-how in managing bilingual data (parallel and comparable)
  - evaluation protocols involving real users (or proxies)
- Better handling data is part of the game
- We should learn from users

Thank you for your attention