

Microblogs as Parallel Corpora

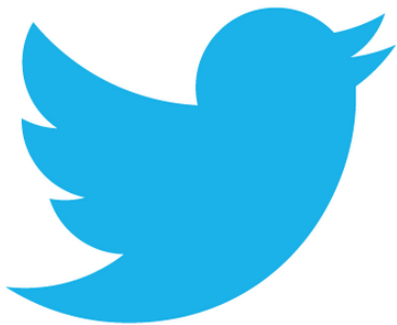
Wang Ling, Guang Xiang,
Chris Dyer, Isabel Trancoso, Alan W Black

Carnegie Mellon University
Instituto Superior Tecnico

In this talk we will...

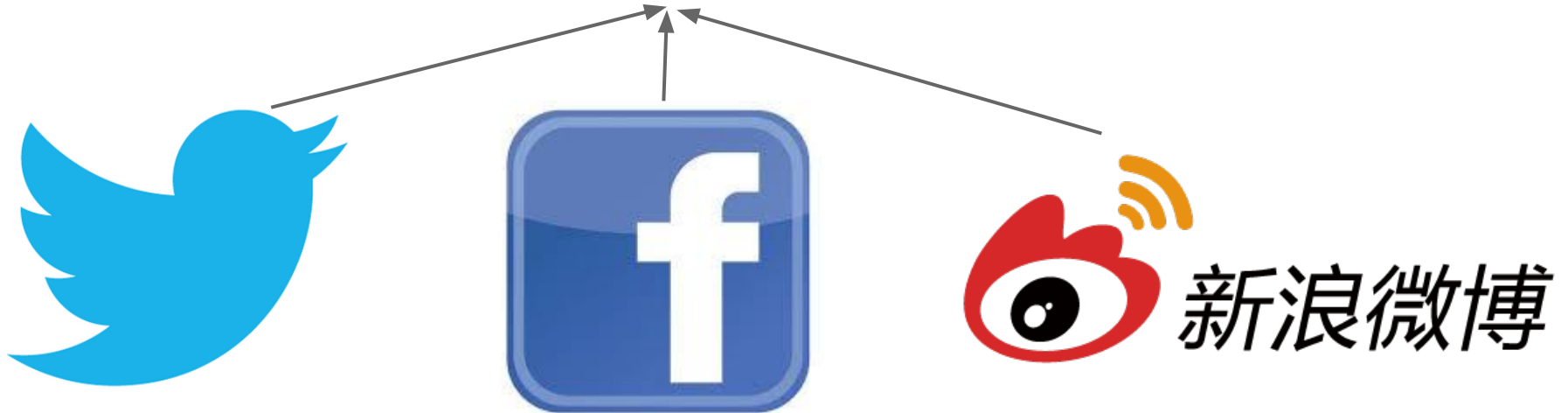
In this talk we will...

- Crawl large amounts of **microblog parallel data for free**



In this talk we will...

- Crawl large amounts of **microblog parallel data for free**

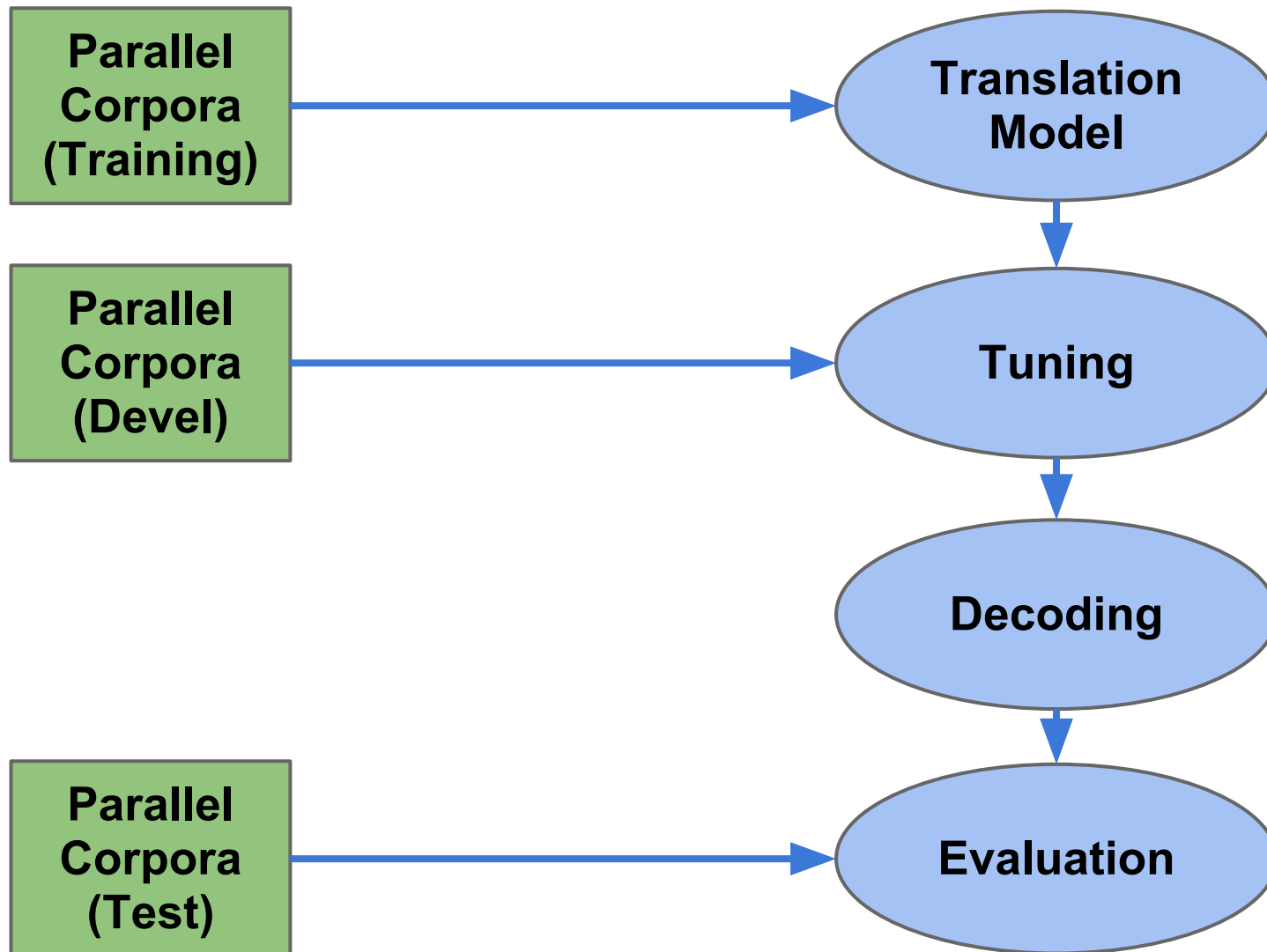


In this talk we will...

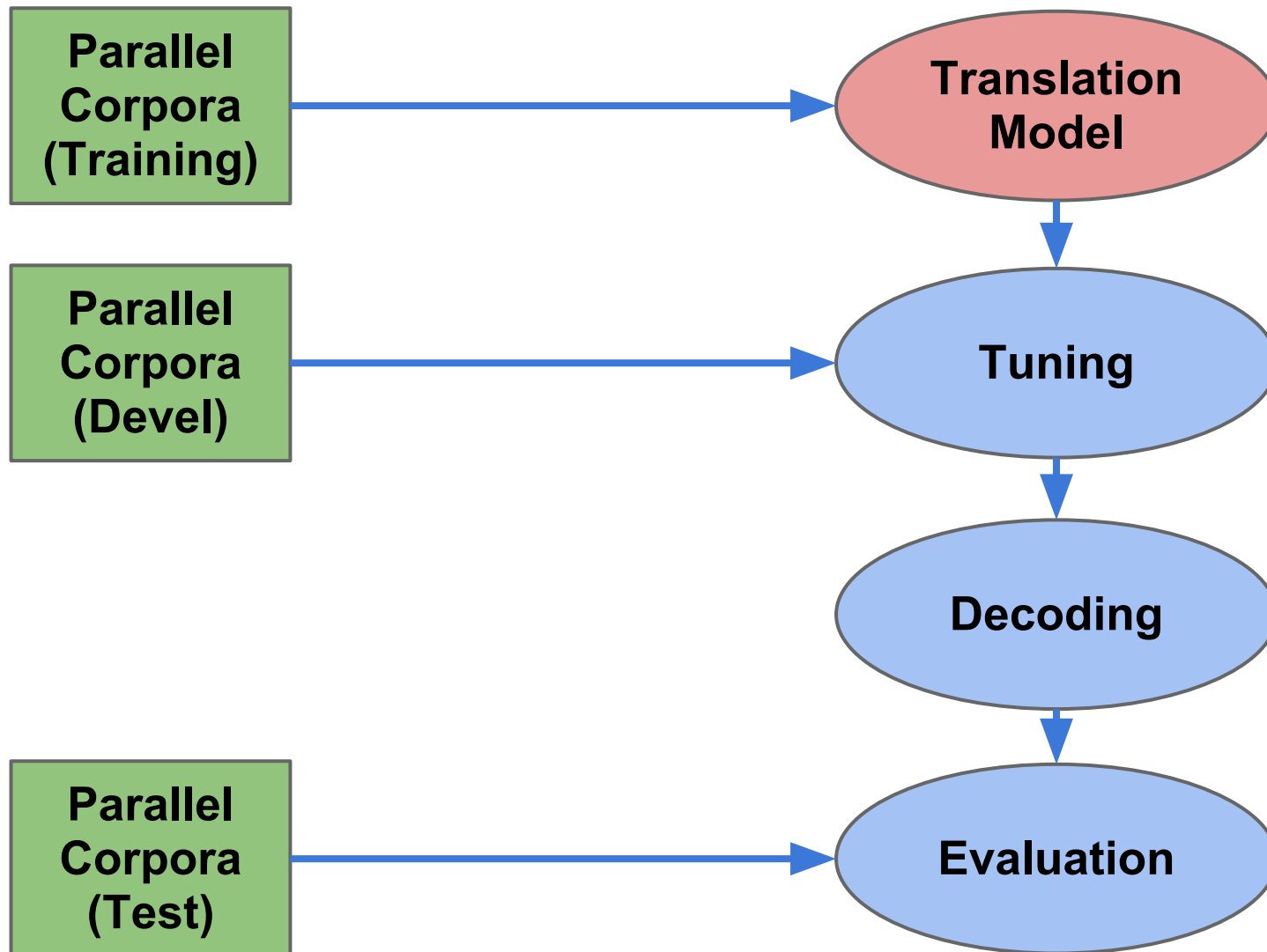
- Crawl large amounts of **microblog parallel data for free**
 - Crawl Sina Weibo (Chinese Twitter)
 - English-Mandarin Pair

Background

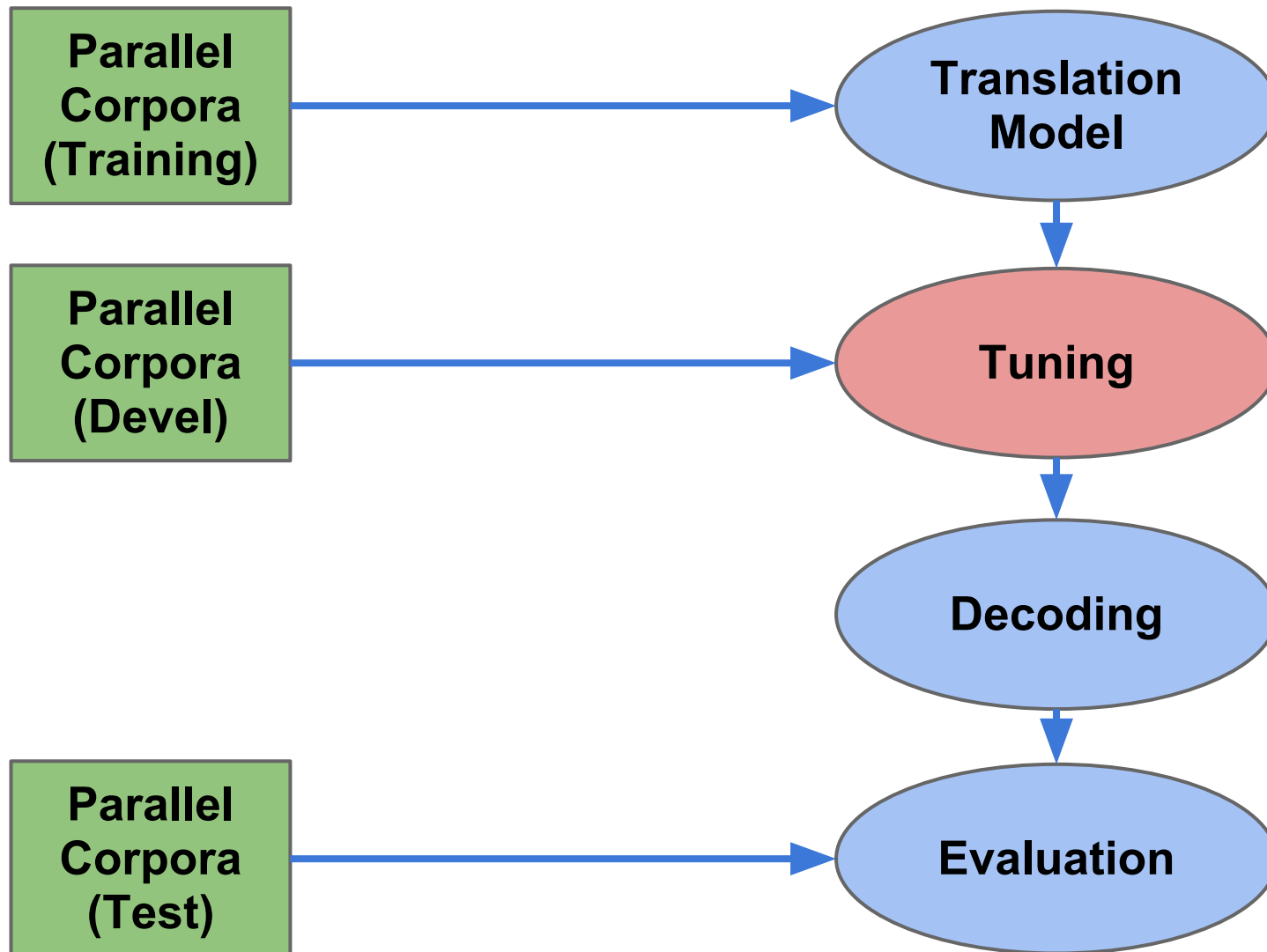
Parallel Data in MT



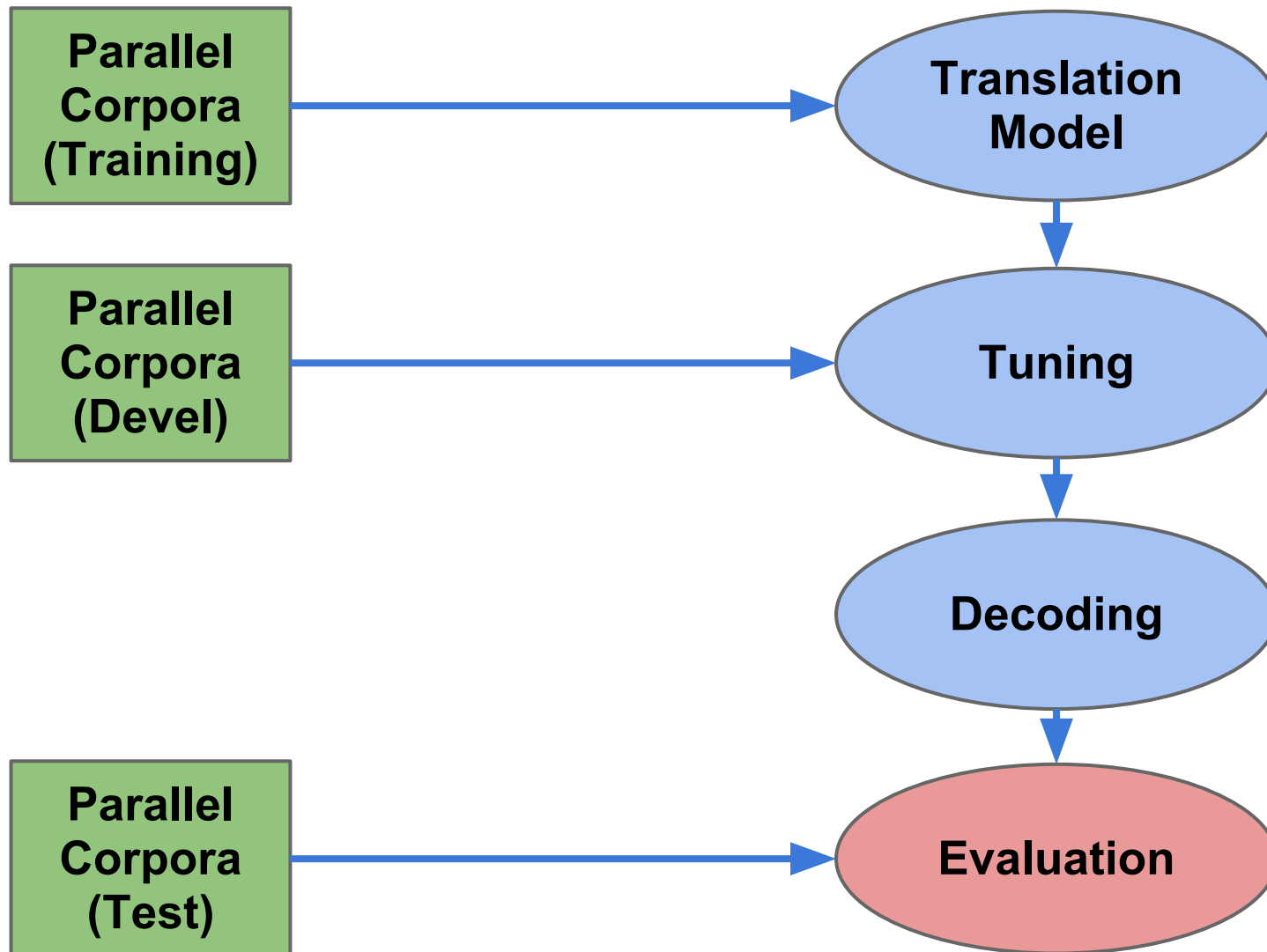
Parallel Data in MT



Parallel Data in MT



Parallel Data in MT



Why do we need Parallel Data from Microblogs?

- Problem: Current parallel corpora are generally **clean and formal**.

MT Model



In 2011, Quebec fell victim to half of the closures and reductions in hours.

Why do we need Parallel Data from Microblogs?

- Problem: Current parallel corpora are generally **clean and formal**. But Microblogs are **noisy and informal**.

MT Model

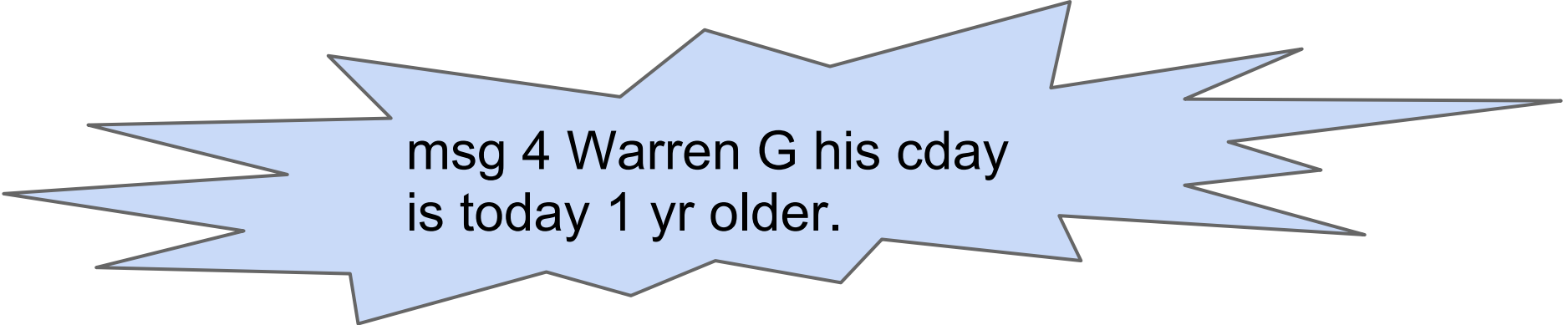


Input



shoutout to the
fans i met today.
love u

Why do we need Parallel Data from Microblogs?



msg 4 Warren G his cday
is today 1 yr older.



Google
Translate

Why do we need Parallel Data from Microblogs?

msg 4 Warren G his cday
is today 1 yr older.

Google
Translate



味精4沃伦G他的cday是今日1年岁。



Why do we need Parallel Data from Microblogs?

msg 4 Warren G his cday
is today 1 yr older.

Google
Translate



味精4沃伦G他的cday是今日1年岁。



Why do we need Parallel Data from Microblogs?

msg 4 Warren G his cday
is today 1 yr older.

Google
Translate



味精4沃伦G他的cday是今日1年岁。



Why do we need Parallel Data from Microblogs?

msg 4 Warren G his **cday**
is today 1 yr older.

Google
Translate



味精4沃伦G他的**cday**是今日1年岁。



Problem with Parallel Data

- Parallel data is a scarce resource



Problem with Parallel Data

- Parallel data is a scarce resource
- Most of the parallel data are crawled from
 - Parallel Websites (Resnik 1999)(Fukushima 2006)
 - Patents (Macken 2007)
 - Parliament data (Koehn 2005)
 - ...



Problem with Parallel Data

- Parallel data is a scarce resource
- Most of the parallel data are crawled from
 - Parallel Websites (Resnik 1999)(Fukushima 2006)
 - Patents (Macken 2007)
 - Parliament data (Koehn 2005)
 - ...
- Crowdsourcing Translation(Zaiden 2011) is an alternative but budget required



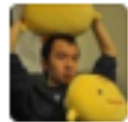
Microblog Parallel Data Extraction

**How can we get Parallel Data in this domain
for free?**



How can we get Parallel Data in this domain for free?

- ...and we found this



Justin Chiu

July 11 near Pittsburgh

I JUST FAILED MY DRIVER LICENCE'S TEST THE FOURTH TIME. EACH TIME I AM WRITING A NEW HISTORY.

考駕照失敗第四次了，我現在真的很好奇我到底要幾次才會過 lol

Like · Comment · Share

Manaal Faruqui, William Yang Wang, Kenton Murray and 19 others like this.



Wang Ling 😊 I can update my presentation now

July 11 at 4:35pm · Like · 2



Justin Chiu Yes, we should keep the most updated result in the publications

July 11 at 4:37pm · Like

Is there Parallel Data in Sina Weibo?

- Does this also happen in Sina Weibo?



Is there Parallel Data in Sina Weibo?

- Does this also happen in Sina Weibo?

Skydiving was incredible! Such an amazing feeling! I loving being adventurous! ;D - 高空 跳伞太不可思议了!真是一种奇妙的感觉 !我喜欢冒险! ;D



Meeting Yao Ming for the first time! So great to be back in China for the Mission Hills World Celebrity Pro-Am. Will post pictures soon! 第一次和姚明见面!又回到中国的感觉太棒了!这次是为观澜湖 世界名人赛。照片稍等片后! Thanks.

Is there Parallel Data in Sina Weibo?

- Formal and Informal

"I am the light and I am the dark. And beyond the light and the dark, I am and God is." 我是 光明, 我也是黑暗。超越光明和黑暗, 我 是, 神是。



msg 4 Warren G his cday is today 1 yr older. happy **cday** may god bless u and the... - 发信息给 Warren G , 今天是他的生日, 又 老了一岁了。生日快乐, 愿上帝保佑你和 ...

Is there Parallel Data in Sina Weibo?

- Formal and Informal

"I am the light and I am the dark. And beyond the light and the dark, I am and God is." 我是 光明, 我也是黑暗。超越光明和黑暗, 我 是, 神是。



msg 4 Warren G his **cday** is today 1 **yr** older. happy **cday** may god bless u and the... - 发信息给 Warren G , 今天是他的生日, 又 老了一岁了。生日快乐, 愿上帝保佑你和 ...

But there is a catch...

But there is a catch...

- **Not all multilingual tweets are parallel**

But there is a catch...

- Not all multilingual tweets are parallel

[GD's Twitter] ONE OF A KIND 的 M/V 马上 就要公开了 !! Y'all Ready for this ? 呃啊 啊啊, 好紧张啊~ 还请大家多多支持 !



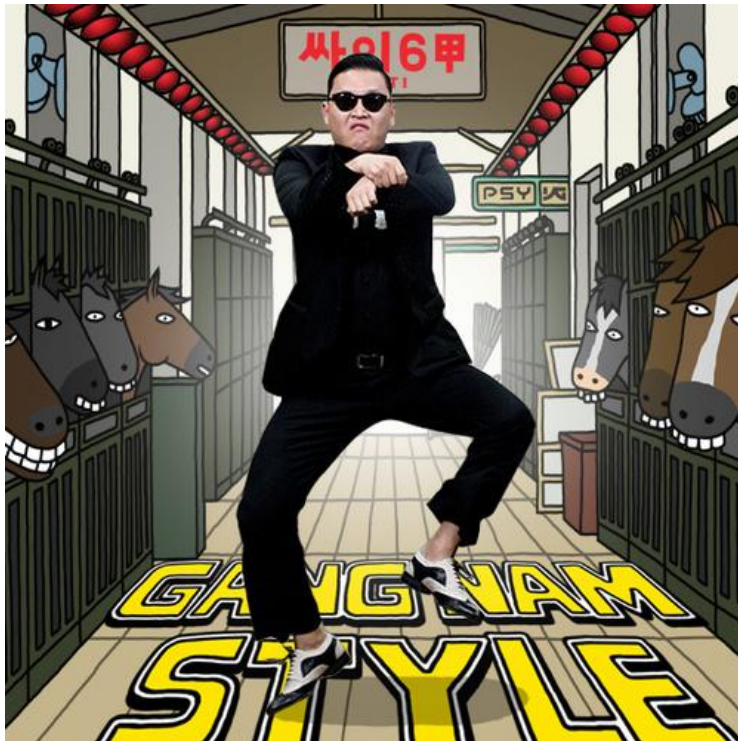
转发微博《南方小羊牧场》 11月9号北 美上映。
Showtime is coming up soon...

But there is a catch...

- Not all multilingual tweets are parallel
- **Finding the parallel segments in the message is not trivial**

But there is a catch...

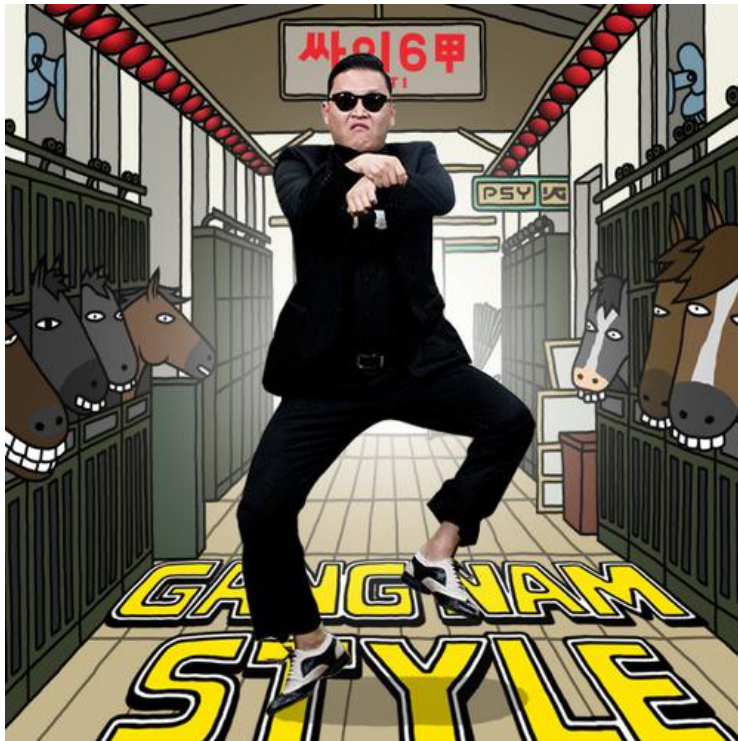
- Not all multilingual tweets are parallel
- Finding the parallel segments in the message is not trivial



I wanna be here every year if possible~! 많은 분들의 걱정처럼 '순간반짝' 일지라도 열심히 해보겠습니다 ... 지나고보면 다 순간이니깐요 ... ^^ **可能的话，我想每年来这里 ~ !** 就算像有的人担心的那样我只是“昙花一现”，我还是会非常努力的 ... 因为回头看的话，一切都只是一瞬的 ... ^^

But there is a catch...

- Not all multilingual tweets are parallel
- Finding the parallel segments in the message is not trivial



I wanna be here every year if possible~! 많은 분들의 걱정처럼 '순간반짝' 일지라도 열심히 해보겠습니다 ... 지나고보면 다 순간이니깐요 ... ^^ 可能的话, 我想每年来这里 ~! 就算像有的人担心的那样我只是“昙花一现”, 我还是会非常努力的 ... 因为回头看的话, 一切都只是一瞬的 ... ^^

Content-based Matching

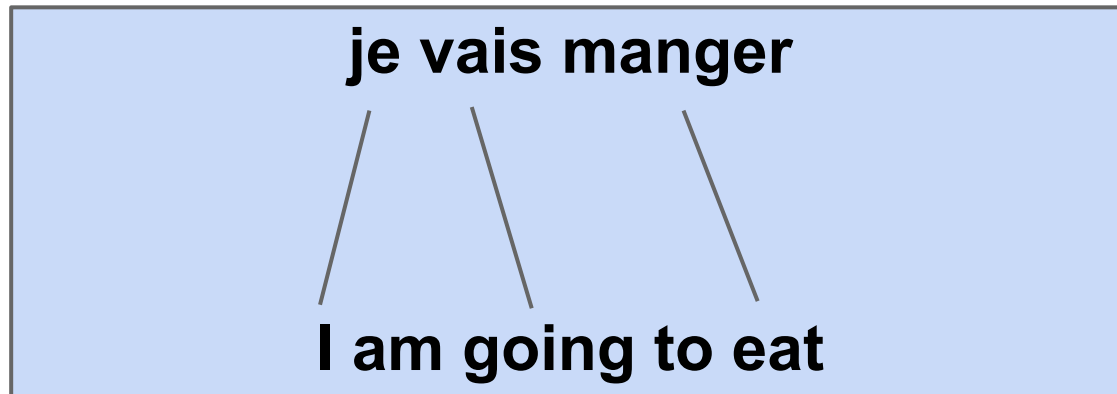
- Given two sentences, calculate their similarity:

je vais manger

I am going to eat

Content-based Matching

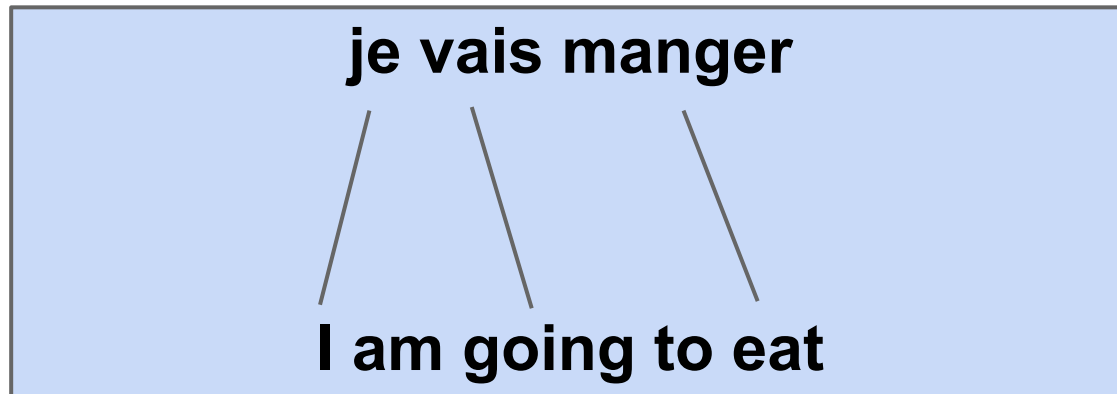
- Given two sentences, calculate their similarity:
 - Compute Viterbi Alignments



Content-based Matching

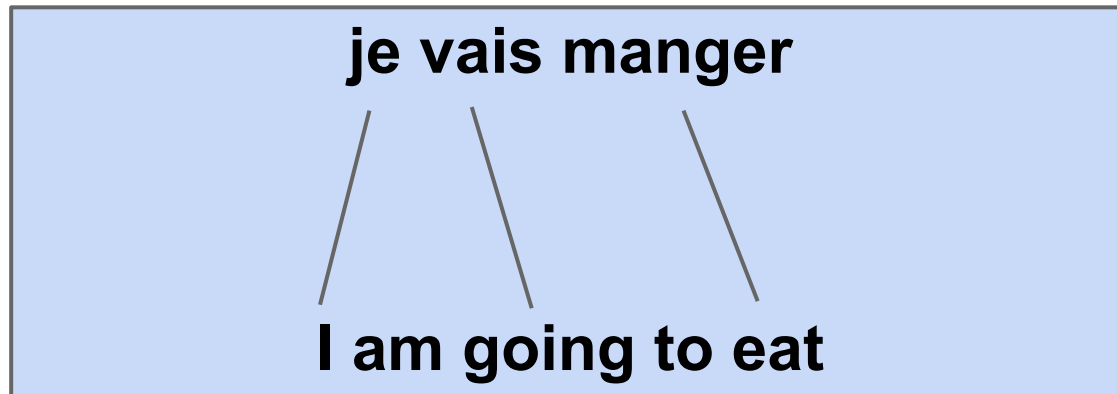
- Given two sentences, calculate their similarity:
 - Compute Viterbi Alignments
 - Compute Similarity Score

$$tsim = \frac{\text{number of alignments}}{\text{number of alignments} + \text{number of unaligned words}} = \frac{3}{5}$$



Content-based Matching

- But, previous work assumes that a pair of documents will be given



Content-based Matching

- ~~But, previous work assumes that a pair of documents will be given~~
- In our case, only one document is provided

je vais manger I am going to eat

Microblog Alignment Model

- Solution: Consider all spans for matching

Microblog Alignment Model

- Solution: Consider all spans for matching

je vais manger I am going to eat

Microblog Alignment Model

- Solution: Consider all spans for matching

je vais manger I am going to eat

Microblog Alignment Model

- Solution: Consider all spans for matching

je vais manger I am going to eat

je vais

going to

Score=0.2

Microblog Alignment Model

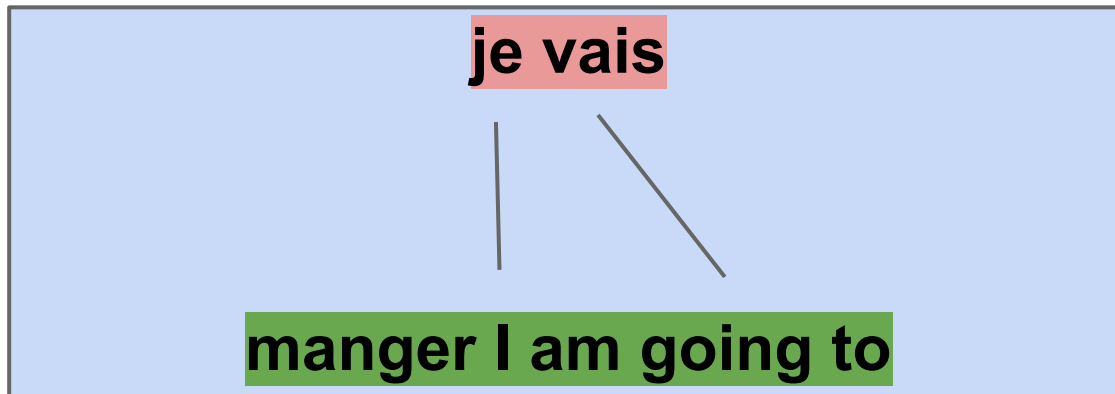
- Solution: Consider all spans for matching

je vais manger I am going to eat

Microblog Alignment Model

- Solution: Consider all spans for matching

je vais manger I am going to eat



Score=0.3

Microblog Alignment Model

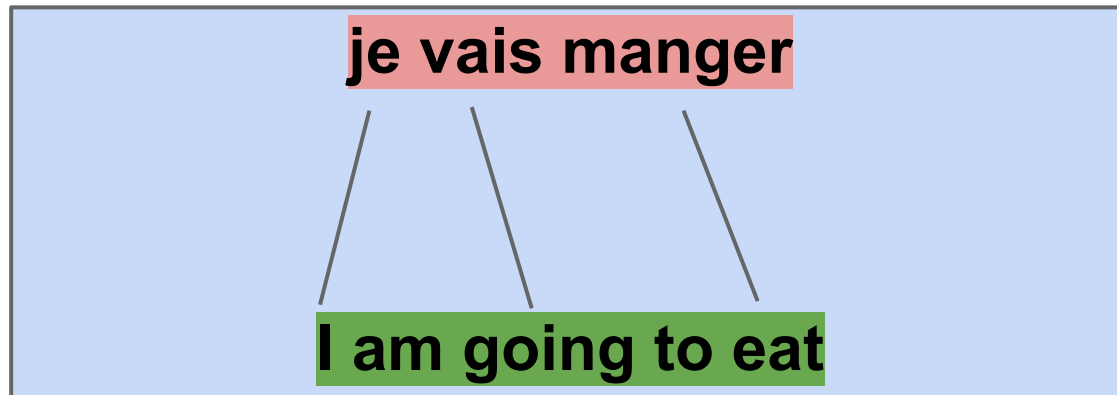
- Solution: Consider all spans for matching

je vais manger I am going to eat

Microblog Alignment Model

- Solution: Consider all spans for matching

je vais manger I am going to eat



Score=0.6

Microblog Alignment Model

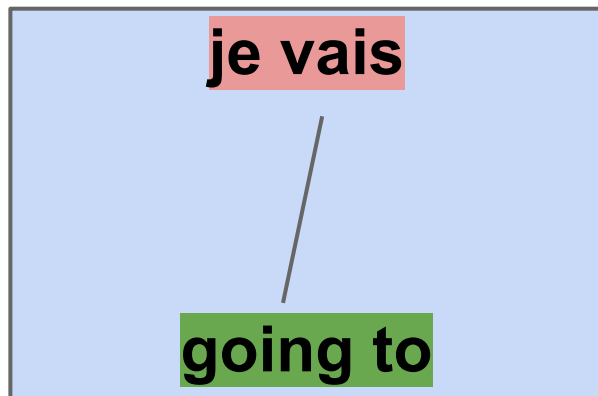
- Solution: Consider all spans for matching
- Problem: Running the Viterbi Alignments for all possible spans is intractable $O(N^6)$:

Microblog Alignment Model

- Solution: Consider all spans for matching
- Problem: Running the Viterbi Alignments for all possible spans is intractable $O(N^6)$:
 - Number of spans = N^4
 - Viterbi alignments = N^2

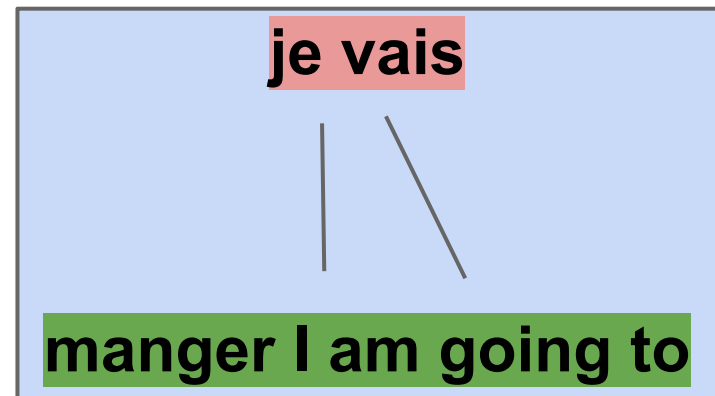
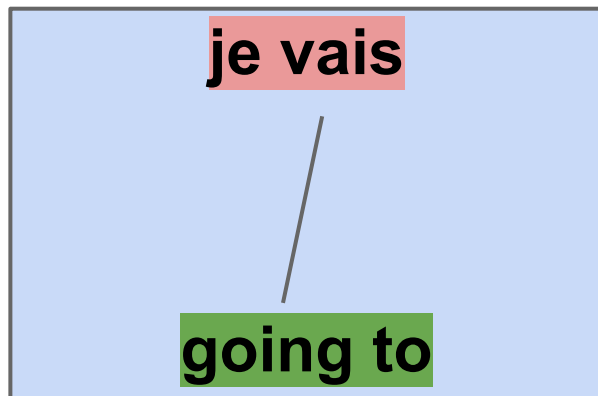
Microblog Alignment Model

- Solution: Consider all spans for matching
- Problem: Running the Viterbi Alignments for all possible spans is intractable $O(N^6)$:
- Answer: Dynamic Programming
 - Reuse Viterbi Alignments for previously processed spans



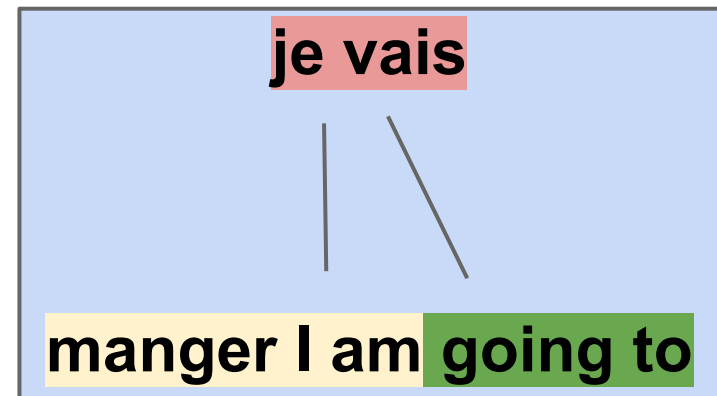
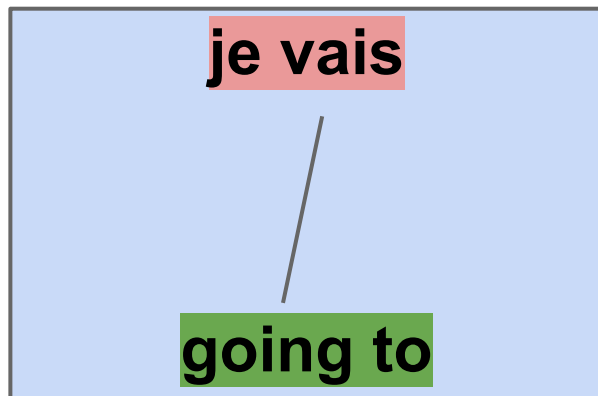
Microblog Alignment Model

- Solution: Consider all spans for matching
- Problem: Running the Viterbi Alignments for all possible spans is intractable $O(N^6)$:
- Answer: Dynamic Programming
 - Reuse Viterbi Alignments for previously processed spans



Microblog Alignment Model

- Solution: Consider all spans for matching
- Problem: Running the Viterbi Alignments for all possible spans is intractable $O(N^6)$:
- Answer: Dynamic Programming
 - Reuse Viterbi Alignments for previously processed spans



Microblog Alignment Model

- Solution: Consider all spans for matching
- Problem: Running the Viterbi Alignments for all possible spans is intractable $O(N^6)$:
- Answer: Dynamic Programming
 - Reuse Viterbi Alignments for previously processed spans
 - Reduces Complexity from $O(N^6)$ to $O(N^4)$

Microblog Alignment Model

- Final score computed by various models

English	Mandarin	Score
You know what?	知道吗？	0.6
You have to remember where you come from b4 u know where u going...	你在知道要去哪里之前先要记得自己从哪里来...	0.5
To DanielVeuleman yea iknw imma work on that	对DanielVeuleman说, 是的, 我知道, 我正在向那方面努力	0.3
just eat it, delicious noodles...	不管多晚, 饿了不吃, 就是睡不着...	0.2

Microblog Alignment Model

- Final score computed by various models
- Extract pairs by thresholding the score

English	Mandarin	Score
You know what?	知道吗？	0.6
You have to remember where you come from b4 u know where u going...	你在知道要去哪里之前先要记得自己从哪里来...	0.5
To DanielVeuleman yea iknw imma work on that	对DanielVeuleman说, 是的, 我知道, 我正在向那方面努力	0.3
just eat it, delicious noodles...	不管多晚, 饿了不吃, 就是睡不着...	0.2

Microblog Alignment Model

- Final score computed by various models
- Extract pairs by thresholding the score

English	Mandarin	Score
You know what?	知道吗？	0.6
You have to remember where you come from b4 u know where u going...	你在知道要去哪里之前先要记得自己从哪里来...	0.5
To DanielVeuleman yea iknw imma work on that	对DanielVeuleman说, 是的, 我知道, 我正在向那方面努力	0.3

Experimental Results

Results

- Dataset
 - Crawled 65 million targeted tweets from Sina Weibo

Results

- Dataset
 - Crawled 65 million targeted tweets from Sina Weibo

Results

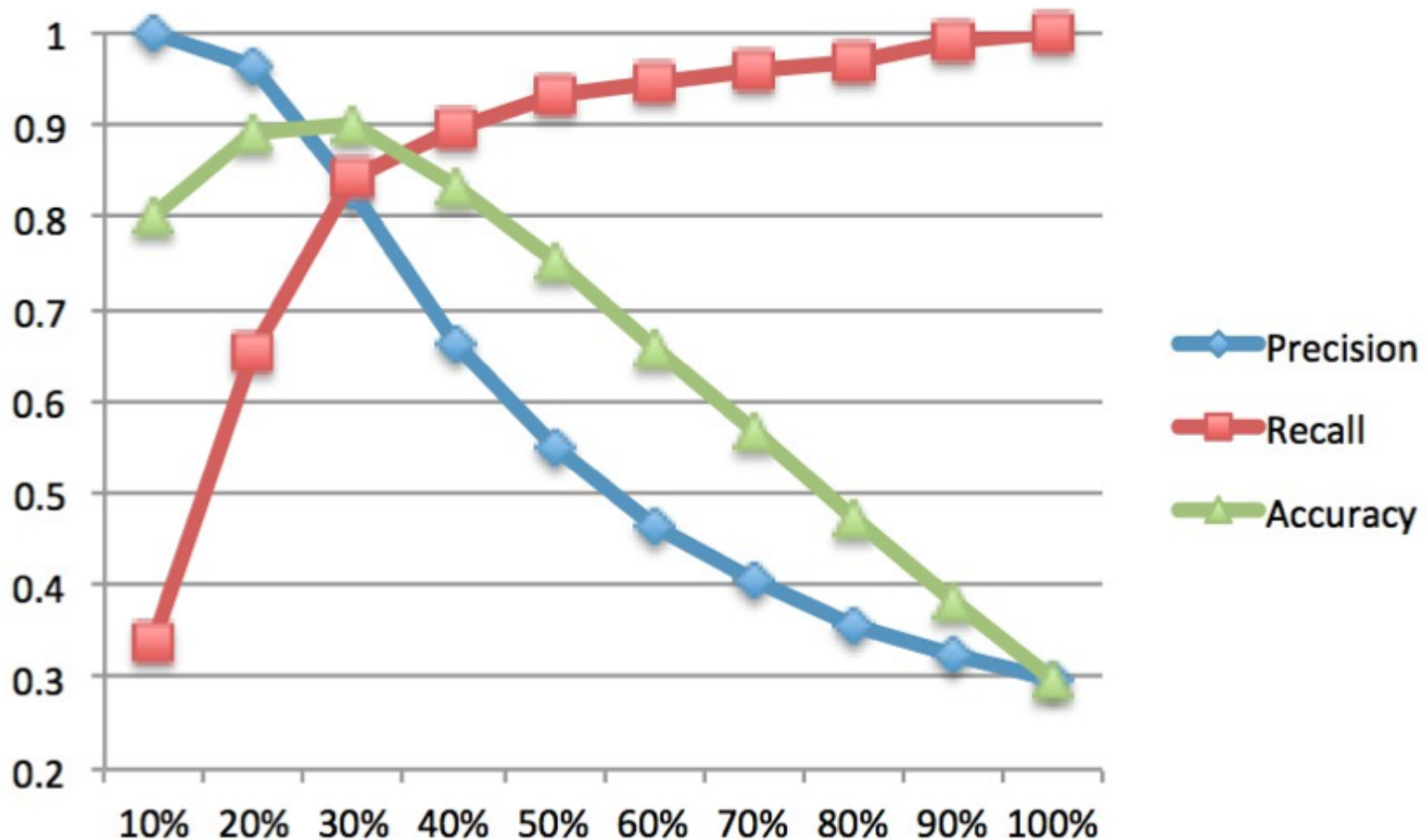
- Dataset
 - Crawled 65 million targeted tweets from Sina Weibo
 - Filtered all tweets with without a Mandarin Trigram and an English Trigram

Parallel Sentence Extraction Results

- Dataset
 - Crawled 65 million targeted tweets from Sina Weibo
 - Filtered all tweets with without a Mandarin Trigram and an English Trigram
- Annotated 2000 tweets sampled uniformly
 - Is the tweet parallel?
 - Where are the parallel spans?

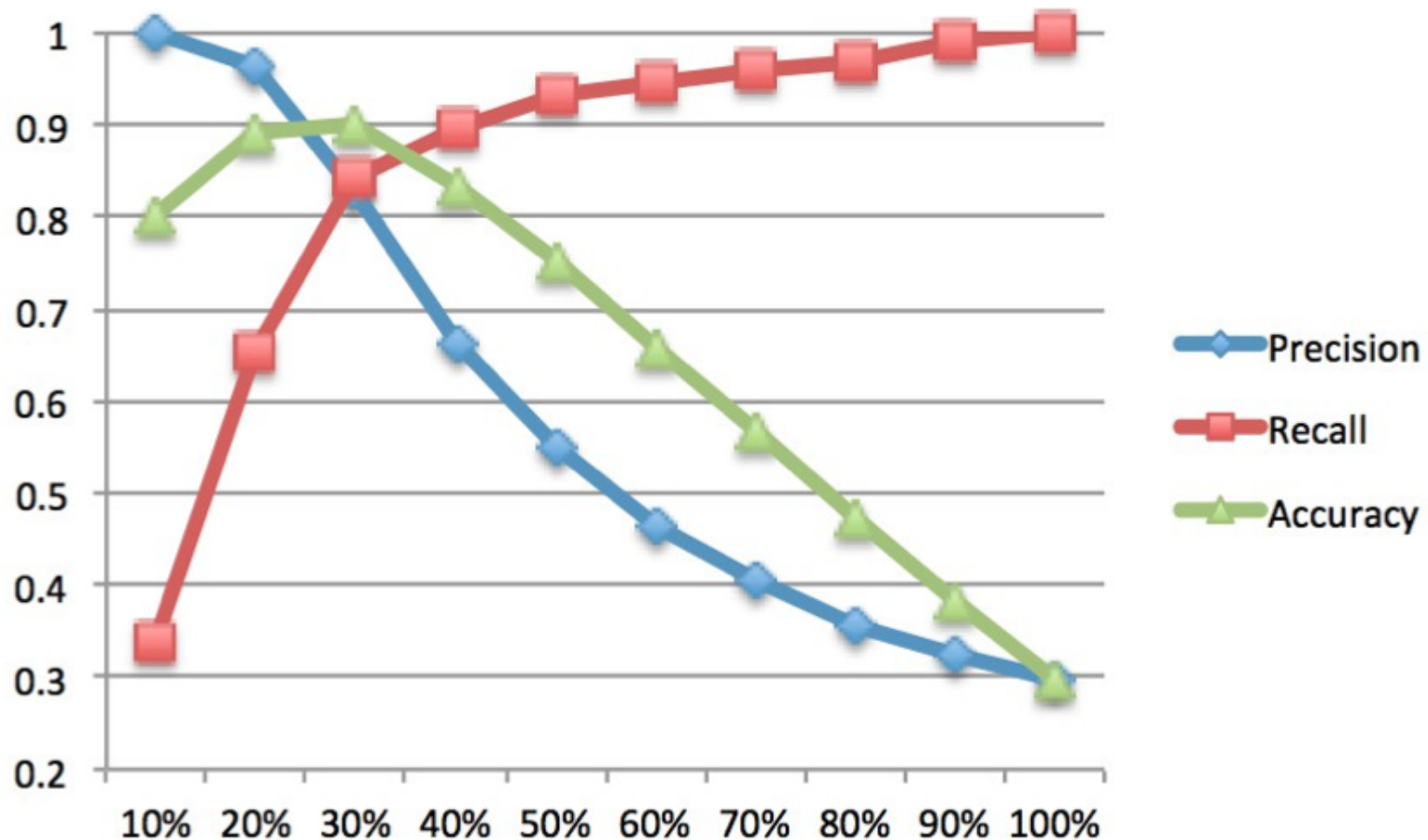
Parallel Sentence Extraction Results

- Parallel Tweet Detection



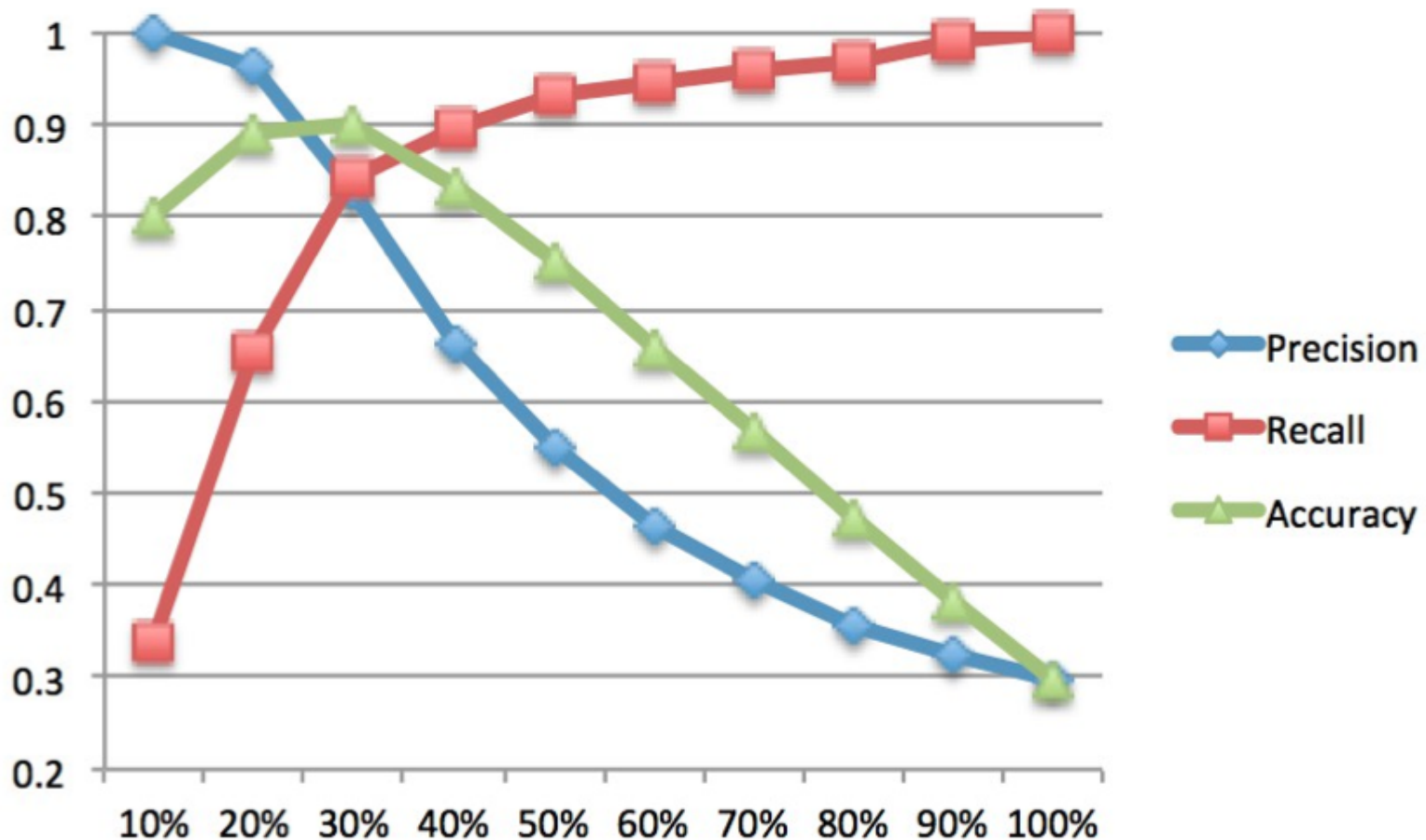
Parallel Sentence Extraction Results

- Keeping 30% of the data is a good trade-off



Parallel Sentence Extraction Results

- 30% of the tweets are parallel



Parallel Sentence Extraction Results

- Span detection:
 - Metric: Average Word Error Rate (no substitutions)

Insertion Error	Deletion Error	Reference
<div><div>je vais manger :D</div><div>I am going to eat</div></div>	<div><div>je vais</div><div>I am going to eat</div></div>	<div><div>je vais manger</div><div>I am going to eat</div></div>

Parallel Sentence Extraction Results

- Span detection:
 - Metric: Average Word Error Rate (no substitutions)

Insertion Error	Deletion Error	Reference
<div><div>je vais manger :D</div><div>I am going to eat</div></div>	<div><div>je vais</div><div>I am going to eat</div></div>	<div><div>je vais manger</div><div>I am going to eat</div></div>




Parallel Sentence Extraction Results

- Span detection:
 - Metric: Average Word Error Rate (no substitutions)

Insertion Error	Deletion Error	Reference
<div><div>je vais manger :D</div><div>I am going to eat</div></div>	<div><div>je vais</div><div>I am going to eat</div></div>	<div><div>je vais manger</div><div>I am going to eat</div></div>

Parallel Sentence Extraction Results

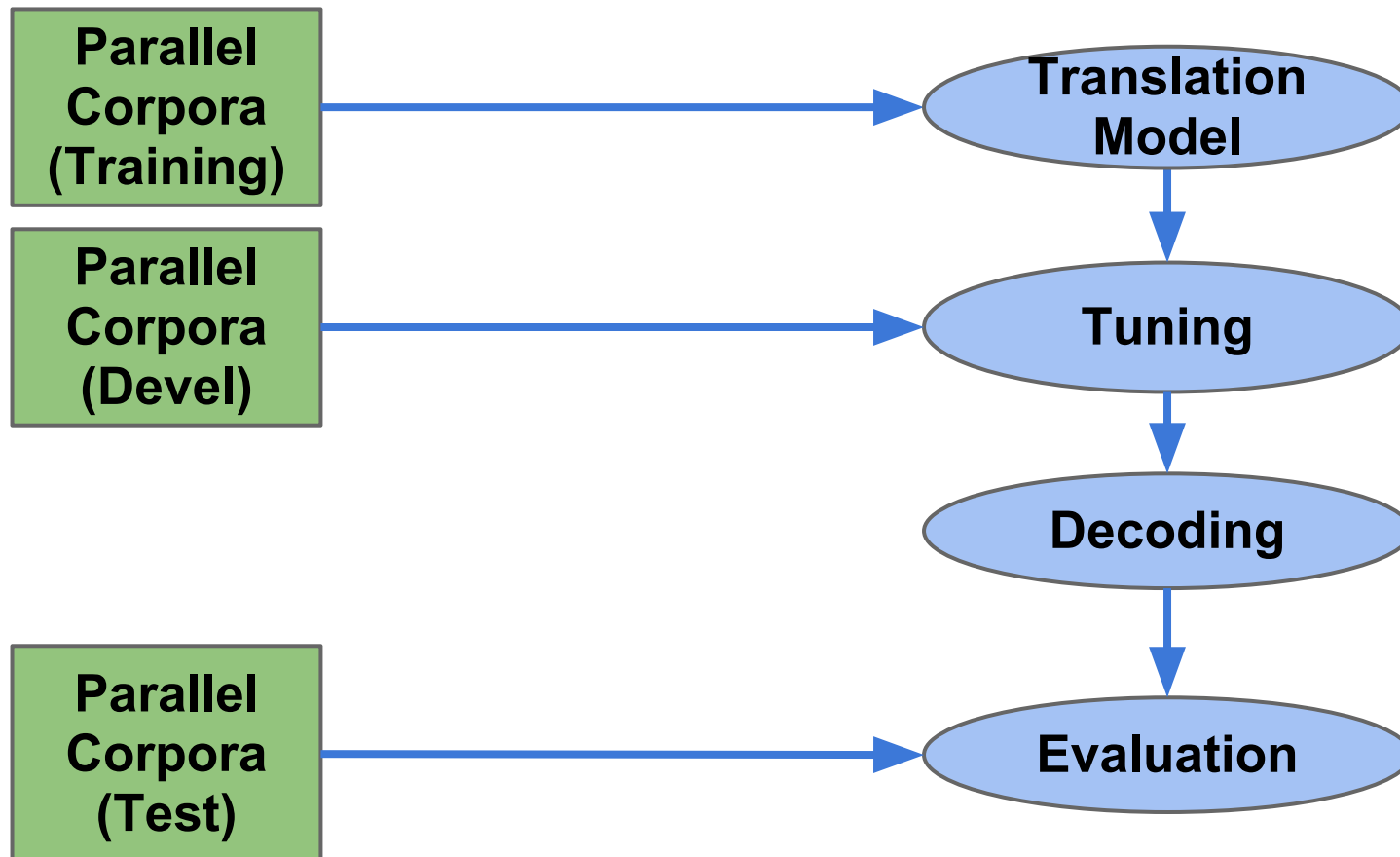
- Span detection:
 - Metric: Average Word Error Rate (no substitutions)
 - WER = 11.4%

Insertion Error	Deletion Error	Reference
		

MT Results

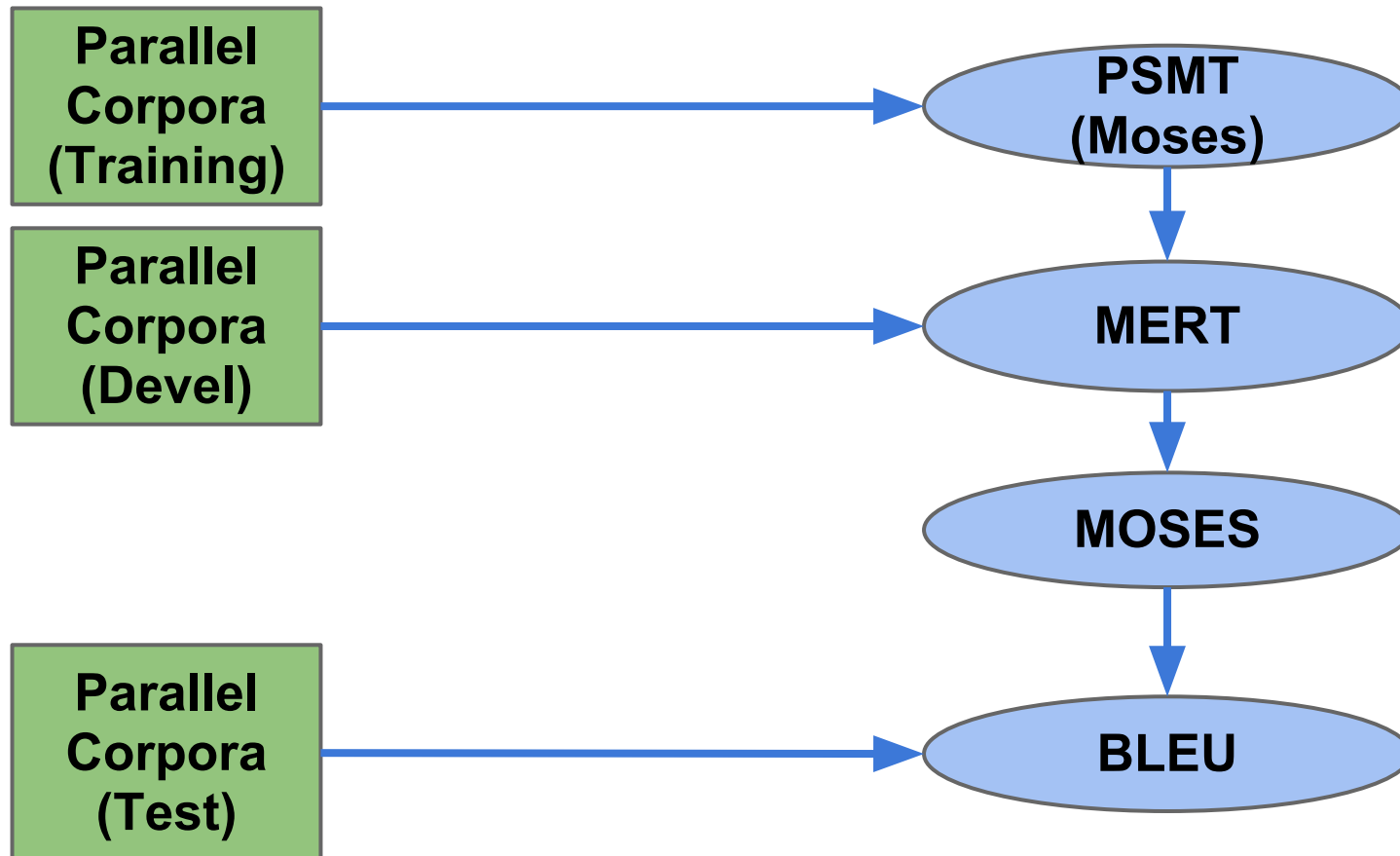
MT Results

- Baseline



MT Results

- Baseline



Results (Extrinsic)

- Training Parallel Data
 - **From Sina Weibo**
 - Approximately 1M multilingual tweets
 - Expect 337K parallel sentences
 - Microblog Domain

Results (Extrinsic)

- Training Parallel Data
 - From Sina Weibo
 - Approximately 1M multilingual tweets
 - Expect 337K parallel sentences
 - Microblog Domain
 - **FBIS dataset**
 - 300K parallel sentences
 - News Domain
 - **NIST dataset**
 - 8M parallel sentences (including FBIS)
 - News Domain

Results (Extrinsic)

- Development and Test sets
 - **Weibo**
 - Built by annotating weibo tweets manually
 - 1000 dev
 - 1000 test
 - Microblog domain

Results (Extrinsic)

- Development and Test sets
 - Weibo
 - Built by annotating weibo tweets manually
 - 1000 dev
 - 1000 test
 - Microblog domain
 - **Syndicate**
 - Extracted from project syndicate (Parallel website)
 - 1000 dev
 - 1000 test
 - News and political domain

Results (Extrinsic)

- MT experiments
 - Significant improvements (30-40%) on microblogs (in-domain)

	Syndicate		Weibo	
	ZH-EN	EN-ZH	ZH-EN	EN-ZH
FBIS	9.4	18.6	10.4	12.3
NIST	11.5	21.2	11.4	13.9
Weibo	8.8	15.9	15.7	17.2

Results (Extrinsic)

- MT experiments
 - Worse results on the Syndicate data(out-of-domain)

	Syndicate		Weibo	
	ZH-EN	EN-ZH	ZH-EN	EN-ZH
FBIS	9.4	18.6	10.4	12.3
NIST	11.5	21.2	11.4	13.9
Weibo	8.8	15.9	15.7	17.2

Results (Extrinsic)

- MT experiments
 - Better results in both datasets by combining parallel data

	Syndicate		Weibo	
	ZH-EN	EN-ZH	ZH-EN	EN-ZH
FBIS	9.4	18.6	10.4	12.3
NIST	11.5	21.2	11.4	13.9
Weibo	8.8	15.9	15.7	17.2
FBIS+Weibo	11.7	19.2	16.5	17.8
NIST+Weibo	13.3	21.5	16.9	17.9

New Translations?

New Translations?

- Abbreviations

谢=thx,你=u

To Colton Lopez, **thx** for the love! 对 Colton Lopez说, 谢
谢你的爱

have u ever really lived in beijing ? 你是否真的住过北京

New Translations?

- Abbreviations

TMD=damn, TM=damn

New Translations?

- Abbreviations

TMD=damn, TM=damn

他妈的—Ta Ma De

New Translations?

- Abbreviations

TMD=damn, TM=damn

Life is like the game "Angry Birds". When you fail, there are always some **damn** stupid pigs laughing at you. 人生就像 "愤怒的小鸟", 当你失败时, 总有 **TMD** 几只笨猪在笑

New Translations?

- Abbreviations
- Jargon

=embarrassed

New Translations?

- Abbreviations
- Jargon

囧=embarrassed

I'm so embarrassed. 我囧死了。

New Translations?

- Abbreviations
- Jargon

囧=embarrassed, 屌丝=loser

New Translations?

- Abbreviations
- Jargon

囧=embarrassed, 屌丝=loser



New Translations?

- Abbreviations
- Jargon

囧=embarrassed, 屌丝=loser

Today I heard a male foreign **loser** roaring in anger on the phone, "You are a liar! You don't love me at all! All you want to do is practise oral English!!! 今天在地铁站, 看到一个外国男**屌丝**在电话咆哮: 你是个骗子! 你一点都不爱我! 你只是想和我练口语!

Related Work

- Jehl et al, 2012, describe a CLIR method to find tweets that are parallel
 - Dataset not available (Tweets cannot be made public)
 - Poster in this ACL (make sure to check it out!)

Conclusion

- Presented an automatic method to extract parallel sentences from microblogs
 - Large amounts of parallel data for free
 - Improvements for the ZH-EN pair

Conclusion

- Presented an automatic method to extract parallel sentences from microblogs
 - Large amounts of parallel data for free
 - Improvements for the ZH-EN pair
- **μtopia - Microblog Translated Posts Corpora**
 - @ <http://www.cs.cmu.edu/~lingwang/microtopia/>
 - 1.5 Million Parallel Sentences from Twitter + Weibo
 - English
 - Mandarin
 - Arabic
 - 7 other languages

Future Work

- **Online Microblog Translation System will be available**
 - @ <http://www.microblogtranslation.org>

Thx y'all 4 ur attention ;)

Thx y'all 4 ur attention ;)

Corpora - <http://www.cs.cmu.edu/~lingwang/microtopia/>

MT system - <http://www.microblogtranslation.org>