An Empirical Comparison of Features and Tuning for Phrase-based Machine Translation

Spence Green

with Daniel Cer and Chris Manning

Stanford University

WMT // 27 June 2014

Recap: ACL13 Results

SGD-based, n-best learning

L_1 feature selection

Recap: ACL13 Results

SGD-based, n-best learning

L_1 feature selection

BOLT-scale Zh−En on NIST data:

	BLEU	Δ
MERT	48.4	

Recap: ACL13 Results

SGD-based, *n*-best learning

L_1 feature selection

BOLT-scale Zh−En on NIST data:

	BLEU	**Δ**
MERT	48.4	
SGD	48.1	

Recap: ACL13 Results

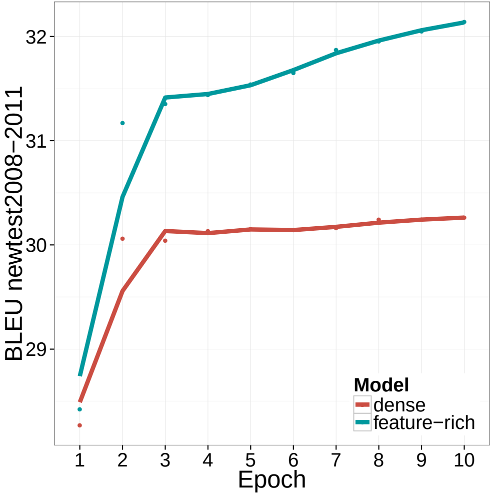
SGD-based, *n*-best learning

L_1 feature selection

BOLT-scale Zh–En on NIST data:

	BLEU	Δ	
MERT	48.4		
SGD	48.1		
SGD+Features	**49.9**	+1.5	:-)

Motivation #1: WMT13 Shared Task :-(



Motivation #1: WMT13 Shared Task

En–Fr news2012 (dev)

	BLEU	
Dense	31.1	
SGD+Features	**31.5**	+0.4

Motivation #2: Practical Issues

Q1: Which **phrase-based features** should I use?

Motivation #2: Practical Issues

Q1: Which **phrase-based features** should I use?

Q2: Why don't my features help?

My Frustrating Summer…

What's wrong with feature-rich MT?

1. Loss Function

My Frustrating Summer…

What's wrong with feature-rich MT?

1. Loss Function
2. References and scoring functions

My Frustrating Summer…

What's wrong with feature-rich MT?

1. Loss Function
2. References and scoring functions
3. Representation: **Features**

My Frustrating Summer…

What's wrong with feature-rich MT?

1. Loss Function
2. References and scoring functions
3. Representation: **Features**

This paper as a pain reliever…

Loss Function

ACL13: Online PRO

Sensitive to length

Doesn’t optimize top-*k*

Slow to compute (sampling)

A Tale about PRO and Monsters

Preslav Nakov, Francisco Guzmán and Stephan Vogel
Qatar Computing Research Institute, Qatar Foundation
Tornado Tower, floor 10, PO box 5825
Doha, Qatar
{pnakov,fherrera,svogel}@qf.org.qa

This work: Online Expected Error

Expected BLEU

$$\ell_t(w_{t-1}) = E_{p_{w_{t-1}}}[-BLEU(d)]$$

$$= -\sum_{d \in H} p_{w_{t-1}}(d) \cdot BLEU(d)$$

This work: Online Expected Error

Expected BLEU

$$\ell_t(w_{t-1}) = E_{p_{w_{t-1}}}[-BLEU(d)]$$

$$= -\sum_{d \in H} p_{w_{t-1}}(d) \cdot BLEU(d)$$

Smooth, non-convex

Fast, less sensitive to length

…but still doesn't prefer top-*k*

References and Scoring

Single vs. Multiple References

Experiment: Compute BLEU+1 for each reference

Single vs. Multiple References

Experiment: Compute BLEU+1 for each reference

Baseline MT system

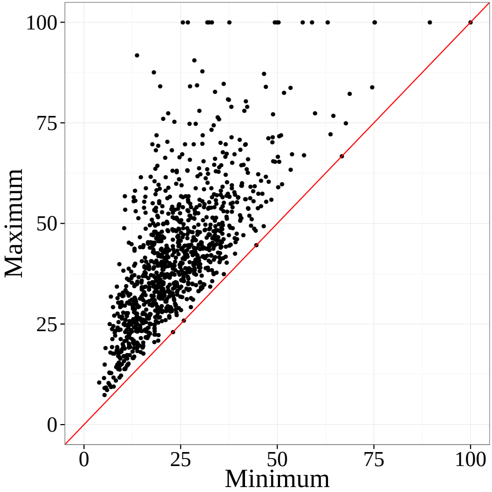
Single vs. Multiple References

Experiment: Compute BLEU+1 for each reference

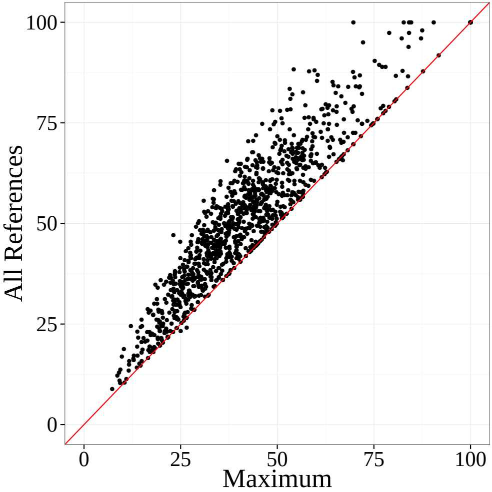
Baseline MT system

Ar–En NIST MT05 has five (5) references

MT05: Max. vs. Min. BLEU+1



MT05: Max. vs. All References BLEU+1



Refs and Scoring Functions

Single-ref Lesson: Don’t try too hard

Refs and Scoring Functions

Single-ref Lesson: Don't try too hard

Blame the **scoring function**?

BLEU+1

BLEU-Nakov [Nakov et al. 2012]

Refs and Scoring Functions

Single-ref Lesson: Don't try too hard

Blame the **scoring function**?

BLEU+1

BLEU-Nakov [Nakov et al. 2012]

BLEU+Noise add Gaussian noise to *n*-gram precisions

Refs and Scoring Functions

Single-ref Lesson: Don't try too hard

Blame the **scoring function**?

BLEU+1

BLEU-Nakov [Nakov et al. 2012]

BLEU+Noise add Gaussian noise to *n*-gram precisions

TER (short translations)

Linear combinations

Representation: Features

Representation: Dense + Extended

Dense features

- Moses baseline templates [Koehn et al. 2007]
- Hierarchical lex. reordering [Galley and Manning 2008]
- Rule count and uniqueness indicator

Representation: Dense + Extended

Dense features

- Moses baseline templates [Koehn et al. 2007]
- Hierarchical lex. reordering [Galley and Manning 2008]
- Rule count and uniqueness indicator

Extended features

Representation: Dense + Extended

Dense features

- Moses baseline templates [Koehn et al. 2007]
- Hierarchical lex. reordering [Galley and Manning 2008]
- Rule count and uniqueness indicator

Extended features

Fire less than dense but more than *sparse*

Goal: a general, robust feature-rich model

Goal: a general, robust feature-rich model

No more ad-hoc features

Goal: a general, robust feature-rich model

No more ad-hoc features

Starting point for more specific features

Five Feature Categories

Common MT error types

1. Lexical Choice
2. Word Alignments
3. Phrase Boundaries
4. Derivation Quality
5. Reordering

Five Feature Categories

Common MT error types

1. Lexical Choice
2. Word Alignments
3. Phrase Boundaries
4. Derivation Quality
5. Reordering

Sources: Novel, literature, word-of-mouth, etc.

Features: Lexical Choice

Filtered Rule Indicator

maison	`maison->the_house`
the house	

Features: Lexical Choice

Filtered Rule Indicator

maison		`maison->the_house`
the house		

Class-based variant

maison		`64->22_14`
the house		

Features: Lexical Choice

Target unigram class

e: utility stocks lead shares higher

Features: Lexical Choice

Target unigram class

e:	utility	stocks	lead	shares	higher
	77	82	3	82	267

Features: Lexical Choice

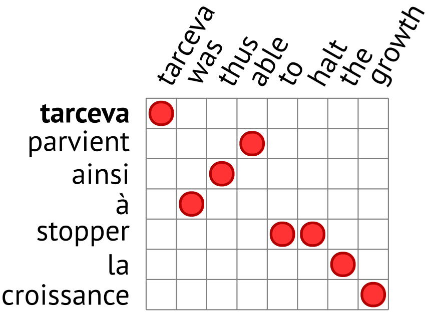
Target unigram class

e:	utility	stocks	lead	shares	higher
	77	82	3	82	267

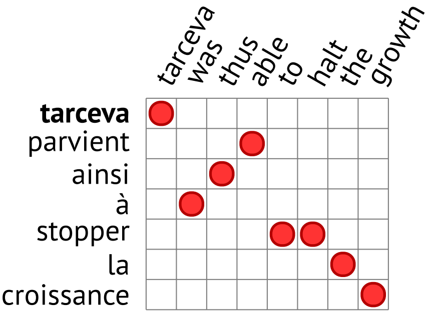
Feature strings:

```
CLASS:77
CLASS:82
CLASS:3
CLASS:82
CLASS:267
```

Features: Word Alignments



Features: Word Alignments



Feature strings:

ALGN:parvient->able

ALGN:stopper->to_halt

etc.

Features: Phrase Boundaries

Target bigram phrase boundary

e: utility | stocks lead shares | higher

Features: Phrase Boundaries

Target bigram phrase boundary

e:	utility	|	stocks	lead	shares	|	higher
	77	|	**82**	3	**82**	|	**267**

Features: Phrase Boundaries

Target bigram phrase boundary

e:	utility	|	stocks	lead	shares	|	higher
	77	|	**82**	3	**82**	|	**267**

Feature strings:

BOUNDARY:77_82

BOUNDARY:82_267

Features: Derivation Quality

Rule dimension features

maison⇒ the house

Features: Derivation Quality

Rule dimension features

maison $\Rightarrow$ the house

Feature strings:

```
SOURCE_DIM:1
TARGET_DIM:2
DIM:1-2
```

Features: Reordering

Filtered Rule Orientation

maison	`SWAP:maison->the_house`
the house	

Features: Reordering

Filtered Rule Orientation

maison the house	SWAP:maison->the_house

Class-based variant

maison the house	SWAP:64->22_14

Aside: Learning Word Classes

Experiment: 3.7M English tokens, 512 classes

Aside: Learning Word Classes

Experiment: 3.7M English tokens, 512 classes

	#threads	**min:sec**
Brown (wcluster)	1	1023:39

Aside: Learning Word Classes

Experiment: 3.7M English tokens, 512 classes

	#threads	**min:sec**
Brown (wcluster)	1	1023:39
Clark (cluster_neyessen)	1	890:11

Aside: Learning Word Classes

Experiment: 3.7M English tokens, 512 classes

	#threads	**min:sec**
Brown (wcluster)	1	1023:39
Clark (cluster_neyessen)	1	890:11
Och (mkcls)	1	199:04

Aside: Learning Word Classes

Experiment: 3.7M English tokens, 512 classes

	#threads	**min:sec**
Brown (wcluster)	1	1023:39
Clark (cluster_neyessen)	1	890:11
Och (mkcls)	1	199:04
This paper	8	**2:42**

[Whittaker and Woodland 2001][Uszkoreit and Brants 2008]

Experiments

NIST Experiments

Stanford Phrasal [Green et al. 2014]

BOLT-scale systems: Ar–En, Zh–En

Four references, uncased BLEU-4

NIST Results: Ar–En

	BLEU	Δ
Dense	42.2	

NIST Results: Ar–En

	BLEU	Δ
Dense	42.2	
Dense+ACL13	44.0	+1.8

NIST Results: Ar–En

	BLEU	Δ
Dense	42.2	
Dense+ACL13	44.0	+1.8
Dense+Ext	44.6	+2.4

NIST Results: Ar–En

	BLEU	**Δ**
Dense	42.2	
Dense+ACL13	44.0	+1.8
Dense+Ext	44.6	+2.4
Dense+Ext+**Domain**	**45.0**	**+2.8**

Domain: feature space augmentation [Daumé III 2007]

NIST Results: Ar–En

	BLEU	Δ
Dense	42.2	
Dense+ACL13	44.0	+1.8
Dense+Ext	44.6	+2.4
Dense+Ext+**Domain**	**45.0**	**+2.8**

Domain: feature space augmentation [Daumé III 2007]

Zh–En: +2.0 BLEU

WMT–14 Shared Task

Single reference, uncased BLEU-4

WMT–14 Shared Task

Single reference, uncased BLEU-4

All Fr–En constrained data

Bilingual		Monolingual
#Segments	**#Tokens**	**#Tokens**
36.3M	2.1M	7.2B

WMT–14 Results: Fr–En

	2014 BLEU	**Δ**	
Dense	35.6		
Dense+Ext	**36.0**	+0.4	:-(

WMT–14 Results: Fr–En

	2014 BLEU	Δ	
Dense	35.6		
Dense+Ext	**36.0**	+0.4	:-(

Uncased BLEU: 1st place

Manual eval: 2–4 cluster

Analysis: Single vs. Multiple References

Ar–En MT09 results

	4-ref	Δ	**1-ref**	Δ
Dense	48.0		47.8	
Dense+Ext	50.0	+2.0	48.9	+1.1

General Observations

More expressive models match refs better (duh)

Single-ref condition == overfitting

General Observations

More expressive models match refs better (duh)

- Single-ref condition == overfitting

Sensitivity to tuning set size/content

- **Bitext tuning**

General Observations

More expressive models match refs better (duh)

Single-ref condition == overfitting

Sensitivity to tuning set size/content

Bitext tuning

Ablation isn't very helpful

Approximate search, non-convex

Conclusion and Impact

Baseline feature-rich representation

Domain adaptation

Conclusion and Impact

Baseline feature-rich representation

Domain adaptation

Faster, better online tuning

Conclusion and Impact

Baseline feature-rich representation

Domain adaptation

Faster, better online tuning

Scalable software to implement the features

See new Phrasal release

An Empirical Comparison of Features and Tuning for Phrase-based Machine Translation

Spence Green

with Daniel Cer and Chris Manning

Stanford University

WMT // 27 June 2014