A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output

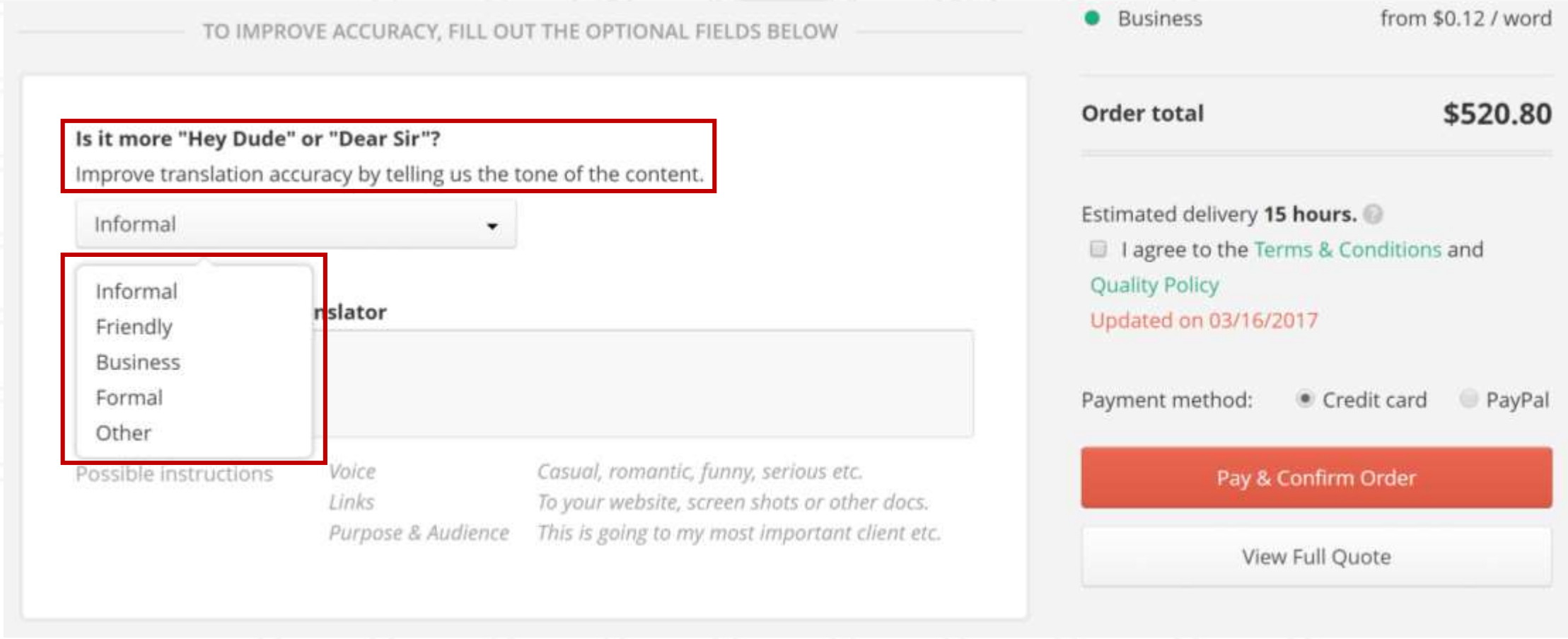
Xing Niu, Marianna Martindale, and Marine Carpuat
CLIP Lab, University of Maryland

DEPARTMENT OF
COMPUTER SCIENCE

2017.09.11 @ EMNLP

- *"Anybody hurt?"*
- *"Is someone wounded?"*

A Human Translation Service*



*gengo.com

Motivation

- *"Anybody hurt?"*
- *"Is someone wounded?"*
  - same literal meaning
  - for different audience or environment

- Goal: controlling formality in machine translation
  - … by asking what is the expected level of formality

  - Prior work looked at other aspects of style
    - politeness in German (Sennrich et al. 2016), gender traits (Rabinovich et al. 2017)

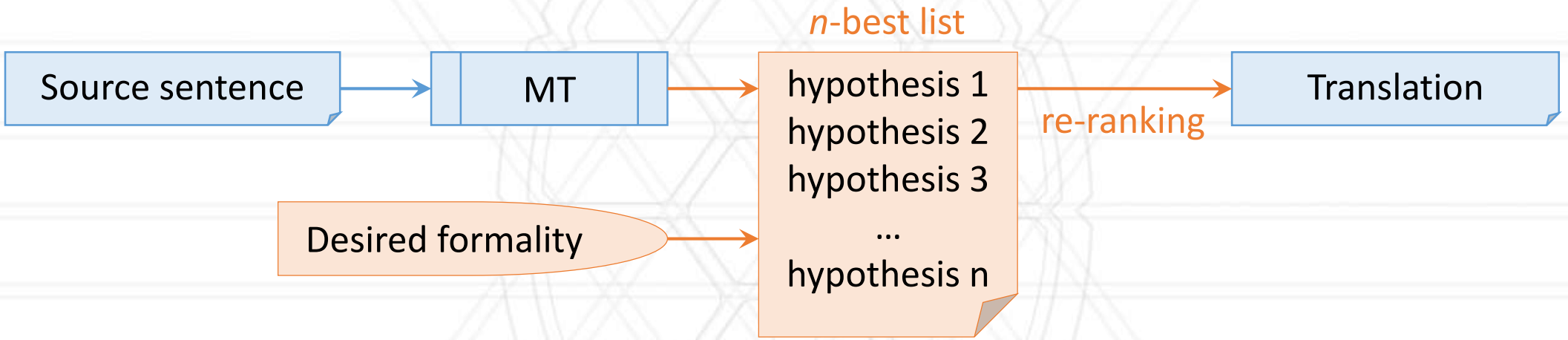
A Study of Style in Machine Translation

- How to control formality in machine translation?
  - Re-ranking-based Formality-Sensitive Machine Translation (FSMT)

Formality-Sensitive MT

Standard MT

Formality-Sensitive MT



- *n*-best list re-ranking with a new feature: $f(h;\ell) = |\,\mathrm{Formality}(h) - \ell|$
  - $\mathrm{Formality}(h)$: the sentence-level formality score [-1,1] of translation hypothesis *h*.
  - ℓ: the desired formality level.

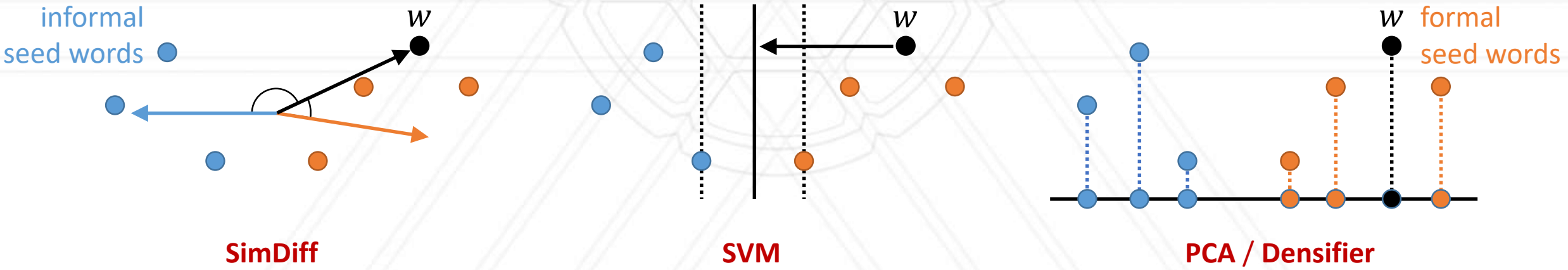
A Study of Style in Machine Translation

- How to control formality in machine translation?
  - Re-ranking-based Formality-Sensitive Machine Translation (FSMT)

- How to score sentential formality?
  - Evaluating existing formality modeling methods

Formality Modeling

- Lexical formality models based on vector space models and formal/informal seed words:
  - **SimDiff** (Brooke et al. 2010) compares words to formal vs. informal seeds.
  - **Support Vector Machine (SVM)** finds a hyperplane that separates seeds.
  - Projecting a word in the high-dimensional space to a one-dimensional score using **Principal Component Analysis (PCA)** or **Densifier** (Rothe et al. 2016).



Formality Modeling – Intrinsic Evaluation

- sentential formality = weighted average of lexical formality
  - Comparing sentential scores with human annotations (11,263 sentences)
    - Lahiri (2015); Pavlick and Tetreault (2016)
  - Metrics: Spearman correlation.

Formality Modeling – Intrinsic Evaluation

	Latent Semantic Analysis (LSA)	Word2vec
SimDiff	0.660	0.654
SVM	0.657	0.585
PCA	0.656	0.663
Densifier	0.664	0.644

Training data for vector space models:
ICWSM 2009 Spinn3r
1.6 billion words from blogs

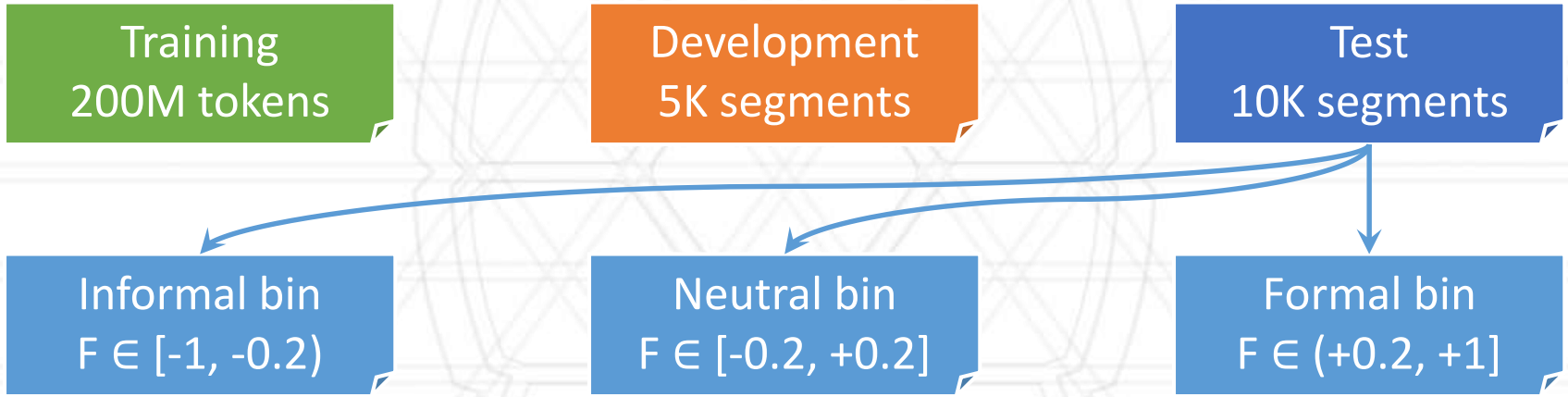
- Models are close in performance.
- Densifier-LSA is selected as a representative for our FSMT system.

A Study of Style in Machine Translation

- How to control formality in machine translation?
  - Re-ranking-based Formality-Sensitive Machine Translation (FSMT)
- How to score sentential formality?
  - Evaluating existing formality modeling methods
- How effective is the FSMT system?
  - Automatic + Human evaluations

Formality-Sensitive MT – Evaluation

- Data: MultiUN + OpenSubtitles2016 (French->English)



- 3 FSMT systems, with different desired formality
  - Low (ℓ=-0.4) | Neutral (ℓ=0) | High (ℓ=+0.4)

FSMT – Automatic Evaluation (BLEU)

Desired formality	Informal test set	Neutral test set	Formal test set
None (baseline)	39.74	40.17	**47.97**
Low	**40.27**	39.65	47.76
Neutral	38.70	**40.46**	47.84
High	37.58	39.53	**47.97**

+0.5

+0.3

Δ=0

Shorter sentence → larger impact.
Formal sentences (MultiUN) are sufficiently different.

- Best: when desired formality level matches reference.

FSMT – Automatic Evaluation (BLEU)

Desired formality	Informal test set	Neutral test set	Formal test set
None (baseline)	39.74	40.17	47.97
Low	**40.27**	39.65	47.76
Neutral	**38.70**	40.46	47.84
High	**37.58**	39.53	47.97

- ΔBLEU $\approx 3 \nrightarrow$ translation quality difference is large.
- BLEU scores conflate translation errors and stylistic mismatch.

FSMT – Human Assessment

- 42 random affected translation pairs for ℓ=±0.4
- 15 volunteers
- Changes in formality:
  - not impact on adequacy
  - small impact on fluency
- According to formality judgment:
  - FSMT impacts 22/42 examples.
  - FSMT correctly yields 21/22 examples w.r.t. formality.
    - more formal output for ℓ=+0.4 than ℓ=-0.4.

FSMT – Human Assessment (Examples)

ℓ	Examples	Comments
-0.4	**anybody hurt** ?	
+0.4	is **someone wounded** ?	annotated as more formal
-0.4	... and then he **ran away** .	
+0.4	... and then he **escaped** .	annotated as more formal
-0.4	he **shot himself** in the middle of it .	
+0.4	he **committed suicide** in the middle of it .	annotated as more formal
-0.4	to move **things** forward .	
+0.4	**in order** to move **the process** forward.	annotated as more formal
-0.4	how do you do ?	annotated as more formal
+0.4	how are you?	

A Study of Style in Machine Translation

- How to control formality in machine translation?
  - Re-ranking-based Formality-Sensitive Machine Translation (FSMT)
- How to score sentential formality?
  - Evaluating existing formality modeling methods
  - Empirical comparison → similar performance
- How effective is the FSMT system?
  - Effective in controlling language formality without loss in translation quality
  - Based on automatic evaluation and human assessment

Code: https://github.com/xingniu/computational-stylistic-variations

DEPARTMENT OF
COMPUTER SCIENCE