



SMS Text Normalization using Statistical Machine Translation

How to translate your daughter's text messages

<http://nlp.stanford.edu/sms/translate.php>

Karthik Raghunathan
 rkarthik@cs.stanford.edu
 Stefan Krawczyk
 stefank@cs.stanford.edu
 Christopher Manning
 manning@stanford.edu

su nlp grp is gr8
 stanford university nlp group is great

OVERVIEW

- Text message normalization is the process of translating 'text speak' into plain English.
- A simple dictionary lookup and replace approach is often unsatisfactory.^{1,2}
 E.g. "do u noe how 2", which one would the dictionary pick: **to, too, two** ?
- Successful translation needs understanding of context and domain knowledge.
 E.g. wishing u a happy new **yr!** -> wishing you a happy new **year!**
 is dat **yr** house? -> is that **your** house?
- We pose this as a statistical machine translation (SMT) task from SMS to English.
- We built a state-of-the-art system using standard machine translation (MT) components.

[1] <http://www.lingo2word.com>
 [2] <http://www.transl8it.com>

challengz

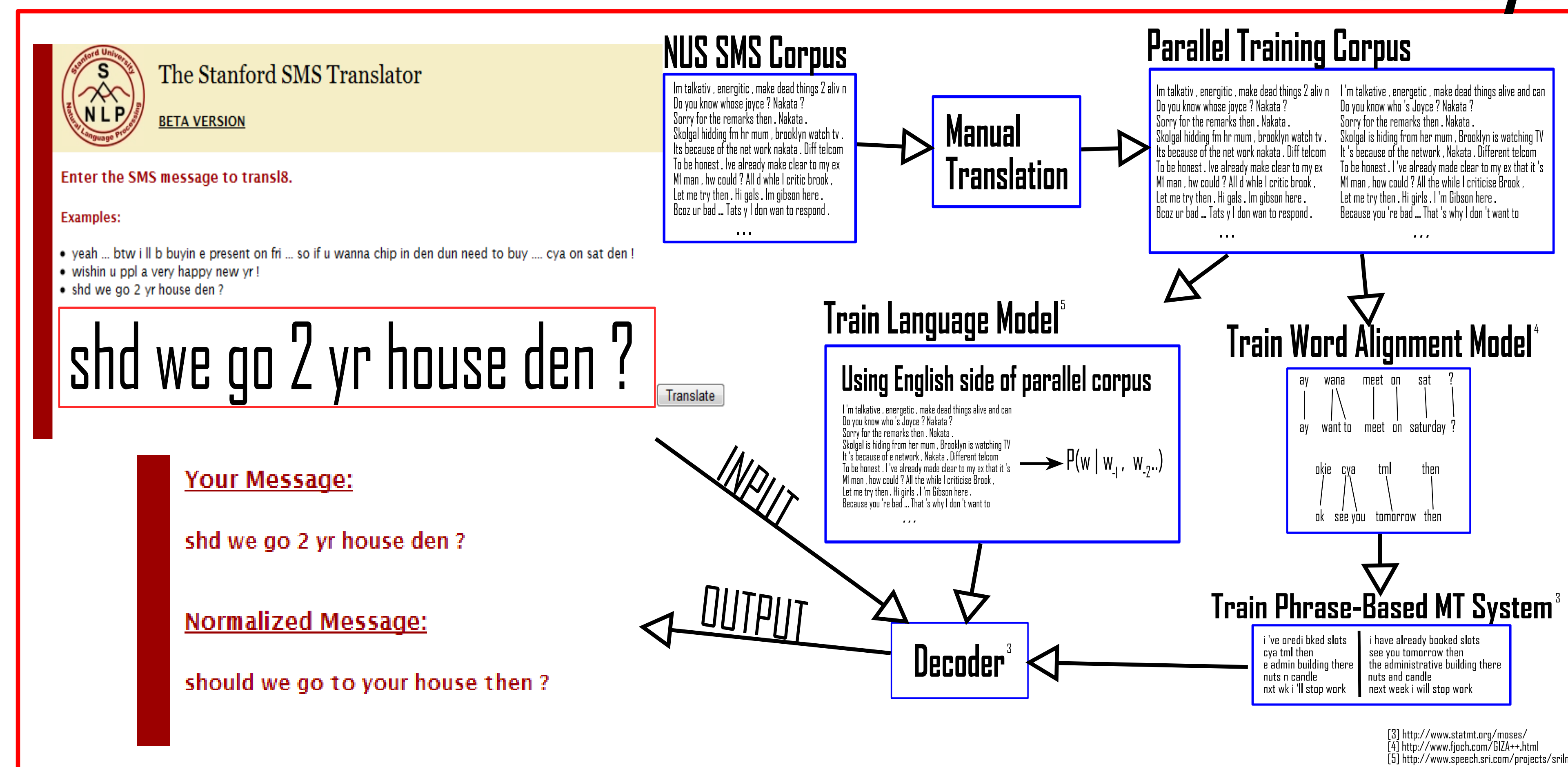
- Disambiguation decisions: 2 -> to, too, two; wk -> week, work.
- To normalize or not? sat -> sat (v.), saturday (n.).
- No formal convention for SMS words: "e" for "the" in Singapore, "da" in the US.
- Colloquial terms: e.g. lea, lor, mah in Singapore.
- Region specific disambiguation: e.g. "mrt" could be "market", but in Singapore refers to a public transport system.
- Topic specific disambiguation: e.g. "asl" in a chatting scenario is "age, sex, language", while in another discussion it could be "American sign language".
- Spelling variations: e.g. wanna -> wana, wannna; people -> ppl, pple.

smt modl

Our model follows the standard phrase-based SMT formulation to translate from text speak to English:

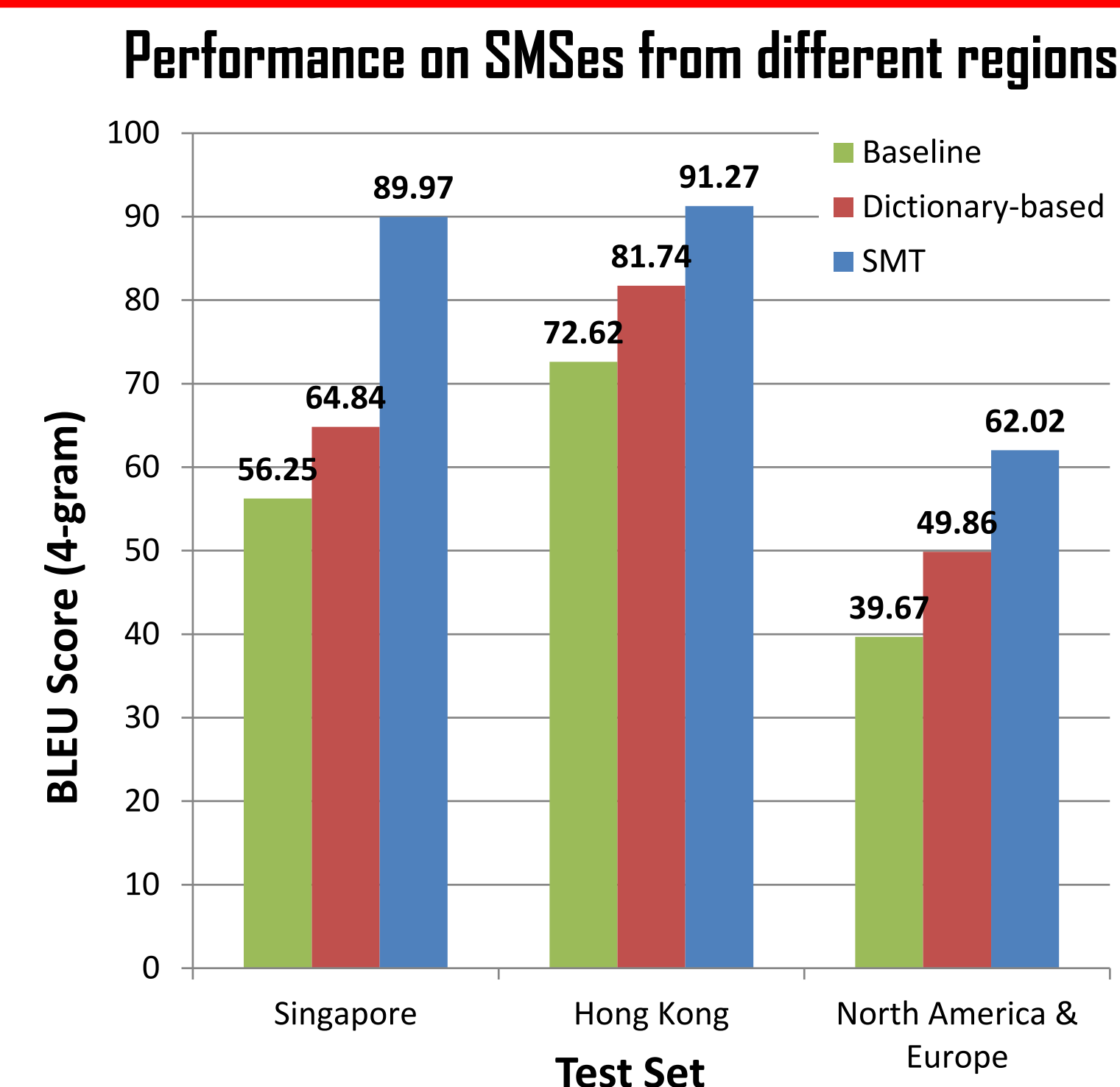
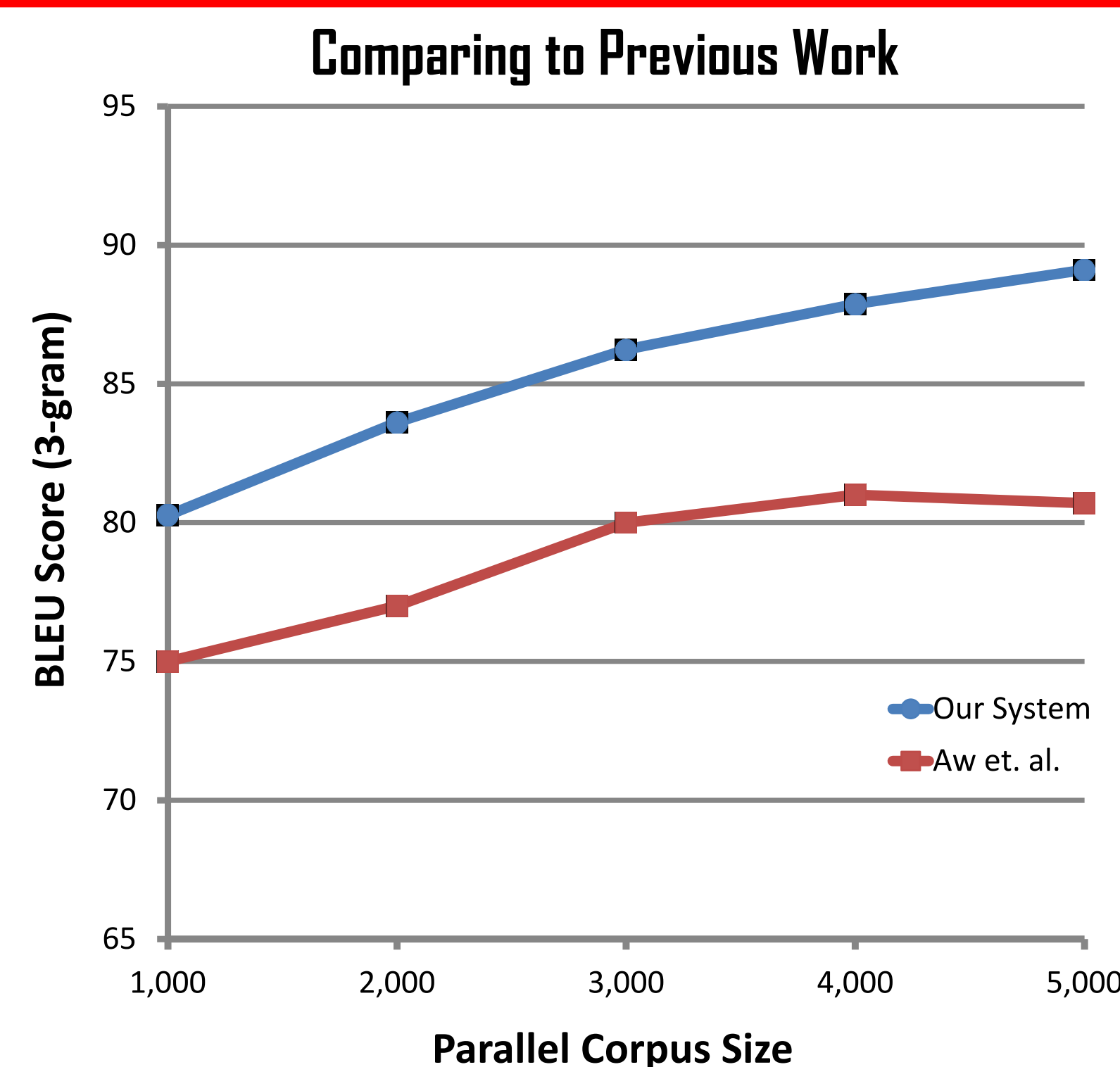
$$E = \underset{E \in \text{English}}{\operatorname{argmax}} \underbrace{\Phi(F|E)}_{\text{phrase translation probability}} \cdot \underbrace{d(F, E)}_{\text{distortion model}} \cdot \underbrace{P(E)}_{\text{language model}} \cdot \underbrace{\omega^{\operatorname{length}(E)}}_{\text{word cost}}$$

smt sys



[3] <http://www.statmt.org/moses/>
 [4] <http://www.fspch.com/BLA.html>
 [5] <http://www.speech.sri.com/projects/serim/>

reslts



Original SMS: yeah ... btw i'll be buyin e present on fri ... so if u wanna chip in den dun need to buy ... cya on sat den !

transl8it.com: yeah ... by the way ill be buyin e present on fri ... so if you want to chip in then dont need to buy see ya on Saturday then !

lingo2word.com: yes sure ... between I will be buying electronic present on Friday ... so if you want to chip in den done need to buy see ya on sat den !

SMT: yeah ... by the way i will be buying the present on friday ... so if you want to chip in then do not need to buysee you on saturday then !

Original SMS: done w tt ages ago .

transl8it.com: done with trailer trash ages ago .

lingo2word.com: done with trailer trash ages ago .

SMT: done with that ages ago .

Original SMS: no lect ah ? i am free from 12 to 2pm .

transl8it.com: know lect ah ? i am free from 12 to 2pm .

lingo2word.com: no lecturer ah ? i am free from dozen to two o'clock in the afternoon

SMT: no lecture ah ? i am free from 12 to 2pm .