Enriching Parallel Corpora for Statistical Machine Translation with Semantic Negation Rephrasing

Dominikus Wetzel[1] Francis Bond[2]

[1]Department of Computational Linguistics
Saarland University
dwetzel@coli.uni-sb.de

[2]Division of Linguistics and Multilingual Studies
Nanyang Technological University
bond@ieee.org

Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation 2012

Untranslated Negations

君 は 僕 に 電話 する 必要 は **ない** 。

$\rightarrow_{reference}$ You need **not** telephone me.

$\rightarrow_{stateOfTheArt}$ You need to call me.

そんな 下劣 な やつ と は 付き合っ て い られ **ない** 。

$\rightarrow_{reference}$ You must **not** keep company with such a mean fellow.

$\rightarrow_{stateOfTheArt}$ Such a mean fellow is good company.

Test data sets	negated	positive
State-of-the-art	22.77	26.60

Table: BLEU for Japanese-English state-of-the-art system.

Distribution of Negations

	Japanese	
English	neg_rel	no neg_rel
neg_rel	8.5%	1.4%
no neg_rel	9.7%	80.4%

- distribution of presence/absence of negation on a semantic level
- Japanese-English parallel Tanaka corpus (ca. 150.000 sentence pairs)
- mixed cases not further explored (lexical negation, idioms)

Method Motivation & Related Work

Suggested method

- produce more samples of phrases with negation
- high quality rephrasing on (deep) semantic structure
- rephrasing introduces new information (as opposed to paraphrasing)
  $\rightarrow$ it needs to be performed on source and target side

- paraphrasing by pivoting in additional bilingual corpora (Callison-Burch et al., 2006)
- paraphrasing with shallow semantic methods (Marton et al., 2009; Gao and Vogel, 2011)
- paraphrasing via deep semantic grammar (Nichols et al., 2010)
- negation handling via reordering (Collins et al., 2005)

Rephrasing Example

	English	Japanese
original	I aim to be a writer.	私 は 作家 を 目指し て いる 。
negations	I don’t aim to be a writer.	私 は 作家 を 目指し て い ない
	I do not aim to be a writer.	私 は 作家 を 目指し て い ませ ん
		私 は 作家 を 目指し ませ ん
		私 は 作家 を 目指さ ない
		作 家 を 私 は 目指し ませ ん
		作 家 を 私 は 目指さ ない

- Japanese: shows more variations in honorification and aspect

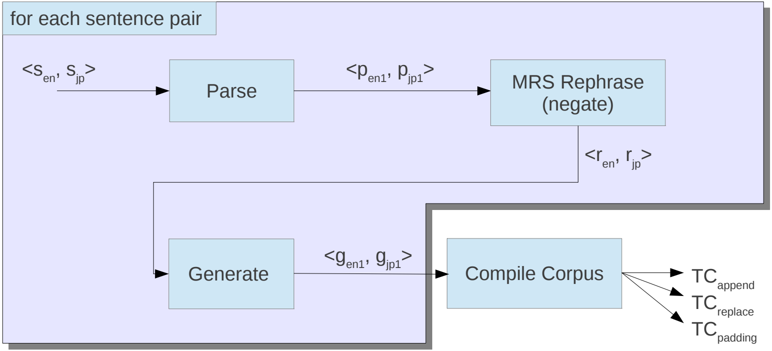
Minimal Recursion Semantics (MRS) – Example

"This may not suit your taste."

$$\begin{bmatrix} \text{TOP} & h1 \\ \text{INDEX} & e2 \\ \text{RELS} & \left\langle \begin{bmatrix} \textit{may\_v\_modal\_rel} & \\ \text{LBL} & h8 \\ \text{ARG0} & e2 \\ \text{ARG1} & h9 \end{bmatrix}, \begin{bmatrix} \textit{neg\_rel} & \\ \text{LBL} & h10 \\ \text{ARG0} & e11 \\ \text{ARG1} & h12 \end{bmatrix}, \begin{bmatrix} \textit{suit\_v\_1\_rel} & \\ \text{LBL} & h13 \\ \text{ARG0} & e14 \\ \text{ARG1} & x4 \\ \text{ARG2} & x15 \end{bmatrix}, \ldots \right\rangle \\ \text{HCONS} & \left\langle h6 =_q h3,\ h12 =_q h8,\ h9 =_q h13,\ \ldots \right\rangle \end{bmatrix}$$

- relevant parts of the English MRS (above)
- necessary parts in the corresponding Japanese MRS are the same

System Overview



Parsing

- bottom-up chart parser for unification-based grammars (i.e. HPSG)
- English Resource grammar (ERG)
  Japanese grammar (Jacy)
- parser, grammar (and generator) from DELPH-IN
- only the MRS structure is required (semantic rephrasing)
- we use the best parse of *n* possible parses for each language; both sides have to have at least one parse
- 84.5% of the input sentence pairs can be parsed successfully

Rephrasing

- add a negation relation EP to the highest scoping predicate in the MRS of each language
- (almost) language abstraction via token identities
- alternatives, where the negation has scope over other EPs are not explored
  more refined changes from positive to negative polarity items are not considered
- 19.6% will not be considered because they are already negated or mixed cases

Generation

- Generator from Lexical Knowledge Builder Environment
- again with ERG and Jacy
- take the highest ranked realization from *n* surface generations of each language; both sides have to have at least one realization
- 13.3% (18,727) of the training data has negated sentence pairs
  $\rightarrow$ mainly because of the brittleness of the Japanese generation

Expanded Parallel Corpus Compilation

- different methods for assembling the expanded version of the parallel corpus (cf. Nichols et al. (2010))
- three versions: Append, Padding and Replace
- use best version also for Language Model (LM) training: Append + negLM

Setup for Japanese-English System

- Moses (phrase-based SMT)
- SRILM toolkit: 5-order model with Kneser-Ney discounting
- Giza++: grow-diag-final-and
- MERT: several tunings for each system (only the best performing ones are considered)

Experiment Data – Token/Sentence Statistics

	Tokens		Sentences	
	train	dev	train	dev
	en / jp	en / jp		
Baseline	1.30 M / 1.64 M	42 k / 53 k	141,147	4,500
Append	1.47 M / 1.84 M	48 k / 59 k	159,874	5,121

- training and development data for SMT experiments: the original Tanaka corpus and our expanded versions

Different Test Sets

Several subsets:
$\rightarrow$ to find out the performance of the baseline and the extended systems on negative sentences

- neg-strict: only negated sentences (based on MRS level)
- pos-strict: only positive sentences (based on MRS level)
- all

Test data sets	all	neg-strict	pos-strict
Sentence counts	4500	285	2684

Results – Japanese-English System

Test data sets	all	neg-strict	pos-strict
Sentence counts	4500	285	2684
Baseline	22.87	22.77	**26.60**
Append	23.01	24.04	26.22
Append + neg LM	**23.03**	**24.40**	26.30

entire test set (all):

- baseline is outperformed by our two best variations Append and Append + neg LM
- differences in BLEU points are 0.14 and 0.16 (not statistically significant)

Results – Japanese-English System

Test data sets	all	neg-strict	pos-strict
Sentence counts	4500	285	2684
Baseline	22.87	22.77	**26.60**
Append	23.01	24.04	26.22
Append + neg LM	**23.03**	**24.40**	26.30

- neg-strict: The gain of our best performing model Append + neg LM compared to the baseline is at 1.63 BLEU points (statistically significant, $p < 0.05$)
- pos-strict: drop of 0.30 and 0.38 in Append + neg LM and Append (both cases statistically insignificant)
- Append + neg LM always performs better than Append

Results – Manual Evaluation of *neg-strict* Test Data

I. decide whether negation is present or not;
quality of translation is not considered:

- systems shown in random order

	Baseline	
Append + neg LM	negation	no negation
negation	51.23%	11.58%
no negation	10.53%	26.67%

Results – Manual Evaluation of *neg-strict* Test Data

II. decide which sentence has a better quality

- systems shown in random order
- score of 0.5 for equal rating
- score of 1 for the better system

Baseline	48.29%
Append + neg LM	51.71%

Discussion

- **baseline**: big decline of performance on neg-strict
  $\rightarrow$ great potential to improve SMT systems by tackling negation problem
- **Append + neg LM**: small decrease on pos-strict, but high increase on neg-strict
  yet, *all* only reflects this high increase to a certain degree
  $\rightarrow$ different proportion of negated and non-negated sentences
- our models are aimed at providing one model which provides a **balance** between this gain and the loss
- providing two **separate** translation models
  $\rightarrow$ direct way to split input data via MRS parsing
  $\rightarrow$ backing-off for undecidable input sentences
- enriched **language model** training data improves BLEU overall; and improves on neg-strict even more

Discussion

- we make use of two existing **large-scale deep semantic grammars**
  $\rightarrow$ more grammars for various languages (German, French, Korean, Modern Greek, Norwegian, Spanish, Portuguese, and more, with varying levels of coverage)
- we **lose input data** along the way: parsing, rephrasing and generation not always successful
  but: twice as many negated pairs in addition; and we do not make use of lower ranked realizations

Conclusion

- alleviates the difficulties of phrase-based SMT with negations $\rightarrow$ problem approached by **expanding the training data** with automatically negated sentence pairs based on **semantic rephrasing**
- small improvements over the baseline considering the entire test data
- performance on negated sentences in the test data shows a statistically significant improvement of 1.63 BLEU points
- also expanding the language model training data boosts performance even more

Future Work

- refine negation rephrasing to have a higher generation rate
- consider more fine grained changes (e.g. negating further embedded predicates, negative polarity items)
- other phenomena could also be tackled in the same way: e.g. rephrasing declarative statements to interrogatives
- combined with the syntactic reordering strategies (Collins et al., 2005) negation reordering rule has more training data $\rightarrow$ a bigger influence on the overall performance
- try out different language pairs (also English–Japanese system); compare low versus high resource settings

References I

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics.

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, Ann Arbor, Michigan. ACL.

References II

Gao, Q. and Vogel, S. (2011). Corpus expansion for statistical machine translation with semantic role label substitution rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 294–298, Portland, Oregon, USA. Association for Computational Linguistics.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Singapore. Association for Computational Linguistics.

Nichols, E., Bond, F., Appling, D. S., and Matsumoto, Y. (2010). Paraphrasing Training Data for Statistical Machine Translation. *Journal of Natural Language Processing*, 17(3):101–122.

Data

- Tanaka corpus (English and Japanese parallel corpus)
- English side: tokenize and truecase
  for evaluation: detruecased and detokenized
- Japanese side: is already tokenized and there are no case distinctions
- Sentences longer than 40 tokens are removed
- baseline: original Tanaka corpus (train: 006-100, dev: 000-002)
- extended corpora: Append, Padding, Replace, Append + neg LM
- train and dev always use the same type of corpus
- test data: profiles 003-005

Background

Minimal Recursion Semantics (MRS)

- top handle, a bag of elementary predicates (EP) and a bag of constraints on handles
- EPs represent verbs, their arguments, negations, quantifiers, etc.
- each EP has a handle with which it can be identified
- top verb introduces an event which is co-indexed with the EP representing the verb

Negation in MRS

- in a negated sentence, the verb being negated is outscoped by the negation relation EP
- a constraint ("equal modulo quantifier") is used to define this scope relation

Distribution of Negations – Mixed Cases

	Japanese	
English	neg_rel	no neg_rel
neg_rel	8.5%	1.4%
no neg_rel	9.7%	80.4%

Table: Distribution of presence/absence of negation on a semantic level.

Mixed cases have two main causes:

- lexical negation such as "She missed the bus." being translated with the equivalent of "She did not catch the bus."
- idioms: such as *ikanakereba naranai* "I must go (lit: go-not-if not-become)" where the Japanese expression of modality includes a negation

Results – Manual Evaluation of *neg-strict* Test Data

II. decide which sentence has a better quality

	Baseline	
Append + neg LM	good	bad
good	28.57%	13.71%
bad	10.29%	47.43%

Expanded Parallel Corpus Compilation

Append

$TC_{append} = \{\}$
for $\langle s_{en}, s_{jp} \rangle \in TC_{original}$ **do**
  $TC_{append} \cup \langle s_{en}, s_{jp} \rangle$
  if hasSuccessfulNegation($\langle s_{en}, s_{jp} \rangle$) **then**
    $TC_{append} \cup \langle negated\ s_{en}, negated\ s_{jp} \rangle$
  end if
end for
return TC_{append}

Expanded Parallel Corpus Compilation

Padding

$TC_{padding} = \{\}$
for $\langle s_{en}, s_{jp} \rangle \in TC_{original}$ **do**
    $TC_{padding} \cup \langle s_{en}, s_{jp} \rangle$
    if hasSuccessfulNegation($\langle s_{en}, s_{jp} \rangle$) **then**
        $TC_{padding} \cup \langle negated\ s_{en}, negated\ s_{jp} \rangle$
    else
        $TC_{padding} \cup \langle s_{en}, s_{jp} \rangle$
    end if
end for
return $TC_{padding}$

- preserving word distribution

Expanded Parallel Corpus Compilation

Replace

$TC_{replace} = \{\}$
for $\langle s_{en}, s_{jp} \rangle \in TC_{original}$ **do**
  if hasSuccessfulNegation($\langle s_{en}, s_{jp} \rangle$) **then**
    $TC_{replace} \cup \langle negated\ s_{en}, negated\ s_{jp} \rangle$
  else
    $TC_{replace} \cup \langle s_{en}, s_{jp} \rangle$
  end if
end for
return $TC_{replace}$

- emphasizing the impact of negated sentences

Results – Japanese-English System

Test data sets	all	biparse	neg-strict	pos-strict	pos-strict-neg-strict
Sentence counts	4500	3399	285	2684	2964
Baseline	22.87	25.76	22.77	**26.60**	26.25
Append	23.01	25.78	24.04	26.22	26.25
Append + neg LM	**23.03**	**25.88**	**24.40**	26.30	**26.28**
Padding	22.74	25.54	22.62	26.35	26.06
Replace	22.55	25.35	23.36	26.00	25.84