The pay-offs of preprocessing for German-English SMT

İlknur Durgar El-Kahlout and François Yvon
LIMSI/CNRS and Université Paris-Sud, Orsay, France

Introduction

- Improving the preprocessing for German-English SMT
  - converting German texts to the new orthographic conventions
  - performing a new tokenization for German
  - normalizing lexical redundancy with the help of POS tagging and morphological analysis
  - splitting German compound words with frequency based algorithm
  - reducing singletons and out-of-vocabulary words
- All processes together **reduce by 10% of the singletons, 2% OOV words**
- **1.5 absolute (7% relative) BLEU improvement** on the WMT 2010 task

SMT with German

- Morphological variation
  - Blur the alignment regularities
  - Increase the level of noise in the phrase table
  - Berg, Berg, Berge, Berges → mountain
- Compound words
  - Composed by concatenating word lemmas
  - No limit on the number of lemmas in a German compound
  - Causes one German word corresponds to several English words, a problematic configuration when aligning German with English

Rinder$_1$kennzeichnungs$_2$- und$_3$ Rindfleisch$_4$ etikettierungs$_5$überwachungs$_6$aufgaben$_7$ übertragungs$_8$ gesetz$_9$

↓

Cattle$_1$ marking$_2$ and$_3$ beef$_4$ labeling$_5$supervision$_6$ duties$_7$ delegation$_8$ law$_9$

  - Morphological productivity also means more out-of-vocabulary (OOV) words, which cannot be translated, at test time

German Preprocessing

Spelling/Orthography Reform

- In 1996, German-speaking countries agreed on an orthography reform (*Rechtschreibreform*)
- To detect old spelling, we used German words *dass* and *muss* and their variants *daß* and *muß*
  - *Sounds and Letters: ß/ss*, *ue/ü*, *ae/ä* and *oe/ö*, Triple Consonants
  - *Foreign Words*: *ph/f*, *gh/g*, *rh/r* and *th/t*.
  - *Use of Hyphens with Numbers:* 400tonner/400-Tonner

Improved Tokenization

- Tokenization is an important, language specific process
- Better tokenization often results in higher translation quality
- *Words With numbers:* 20-Tonner/20 Tonner, 65-mal/65 mal.
- *Words with hyphens:* Getrennt- und Zusammenschreibung /Getrennt und Zusammenschreibung

Compound Splitting

Compound splitting

- Frequency based algorithm (Koehn and Knight, 2003)
- Min. 4-8 characters for split and candidate words
- Fillers:
  Insertion (-s -n -en -nen -e -es -er -ien)
  Truncation (-e -en -n)
- Split mark except for the last word

OOV lemmatization

- Unseen morphological variants
- New words created by "pseudo" tagging
- New words created by compound split marking

Removing Lexical Redundancy

Input	POS	Lemma	Analysis
In	APPR	in	APPR.In
der*	ART	d	ART.Def.Dat.Sg.Fem
Folge	NN	Folge	N.Reg.Dat.Sg.Fem
befand	VVFIN	befinden	VFIN.Full.3.Sg.Past.Ind
die*	ART	d	ART.Def.Nom.Sg.Fem
derart	ADV	derart	ADV
gestÄärkte*	ADJA	gestÄärkt	ADJA.Pos.Nom.Sg.Fem
Justiz	NN	Justiz	N.Reg.Nom.Sg.Fem
wiederholt	ADJD	wiederholt	ADJD.Pos
gegen	APPR	gegen	APPR.Acc
die*	ART	d	ART.Def.Acc.Sg.Fem
Regierung	NN	Regierung	N.Reg.Acc.Sg.Fem
und	KON	und	CONJ.Coord.-2
insbesondere	ADV	insbesondere	ADV
gegen	APPR	gegen	APPR.Acc
deren*	PDAT	d	PRO.Dem.Subst.-3.Gen.Sg.Fem
Geheimdienste*	NN	Geheimdienst	N.Reg.Acc.Pl.Masc
.	$.	.	SYM.Pun.Sent

TreeTagger and RFTagger outputs

Some typical rules :

- **For articles, adjectives (only positive form) and pronouns (indefinite , possessive, demonstrative and relative pronouns);**
  - If a token has genitive case: replace with lemma+en
  - If a token has plural number: replace with lemma+s
  - All other gender, case and number: replace with lemma
- **For nouns;**
  - Plural number: replace with lemma+s
  - All other gender and case: replace with lemma
- **For main verbs (except auxiliary and modal verbs);**
  - The verbs with present tense and 3rd person singular is not changed.
  - All other verbs: lemma+past or lemma+pres, depending on the tense

Experimental Settings

- Data: All available WMT10 texts
- Training: *GIZA++* and *grow-diag-final-and*
- Tuning: *MERT*
- LM: Standard 4-gram LM was estimated on separate parts with Knesser-Ney discounting
- LMs are linearly interpolated (coefficients are selected to minimize the perplexity)

Results

- Both normalization and compound splitting help to increase the translation quality
- To see the effect of the combination of these two methods, we splitted the compounds of the best normalization configuration
- We used 4-8 as candidate-split minimum character lengths with only addition suffixes

System	*BLEU*
Baseline	20.03
Spelling Refom	20.45
+New Tokenization	20.55
+Normalization	20.85
+Compound Splitting	21.27
+Singleton Normalization	21.35
+OOV Normalization	**21.46**

Results on WMT 2010 newstest2010 test data

ART	*ADJ*	*PRO*	*NOUN*	*VERB*	*BLEU*
+					20.76
	+				20.75
		+			20.65
			+		20.64
				+	20.54
	+		+		20.75
+			+		20.54
+	+		+		20.49
	+	+	+		**20.85**
	+		+	+	20.32
	+	+	+	+	20.48
+	+	+	+		20.67
+	+	+	+	+	20.69

German-English normalization results in detail

System	*Unique W.*	*Singletons*	*OOV*
Spelling Reform	382k	50.4%	5.2%
New Tokenization	374k	49.9%	5.1%
Normalization	368k	51.9%	5.0%
Compound Splitting	220k	45.4%	3.4%
Singleton Norm.	205k	40.2%	3.6%
OOV Norm.	–	–	3.2%

German Statistics for various experiments