A World Leading SFI Research Centre

Science Foundation Ireland For what's next

Engaging Content
Engaging People

Maintaining sentiment polarity in translation of user-generated content

Pintu Lohar, Haithem Afli and Andy Way

ADAPT Centre, School of Computing, Dublin City University

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Contents

Objective

- Analyse sentiment preservation & MT quality in the context of user-generated content (UGC)

Objective

- Analyse sentiment preservation & MT quality in the context of user-generated content (UGC)

- Focus on whether sentiment classification helps improve sentiment preservation in MT of UGC

Motivation

- Translation quality *per se* is not the main concern

Motivation

- Translation quality *per se* is not the main concern

- **Sentiment preservation** is (arguably) more important

e.g. companies want to know what their customers think of their products and services.

It is **crucial** that user sentiment in one language is preserved in the target language (typically, English).

Motivation

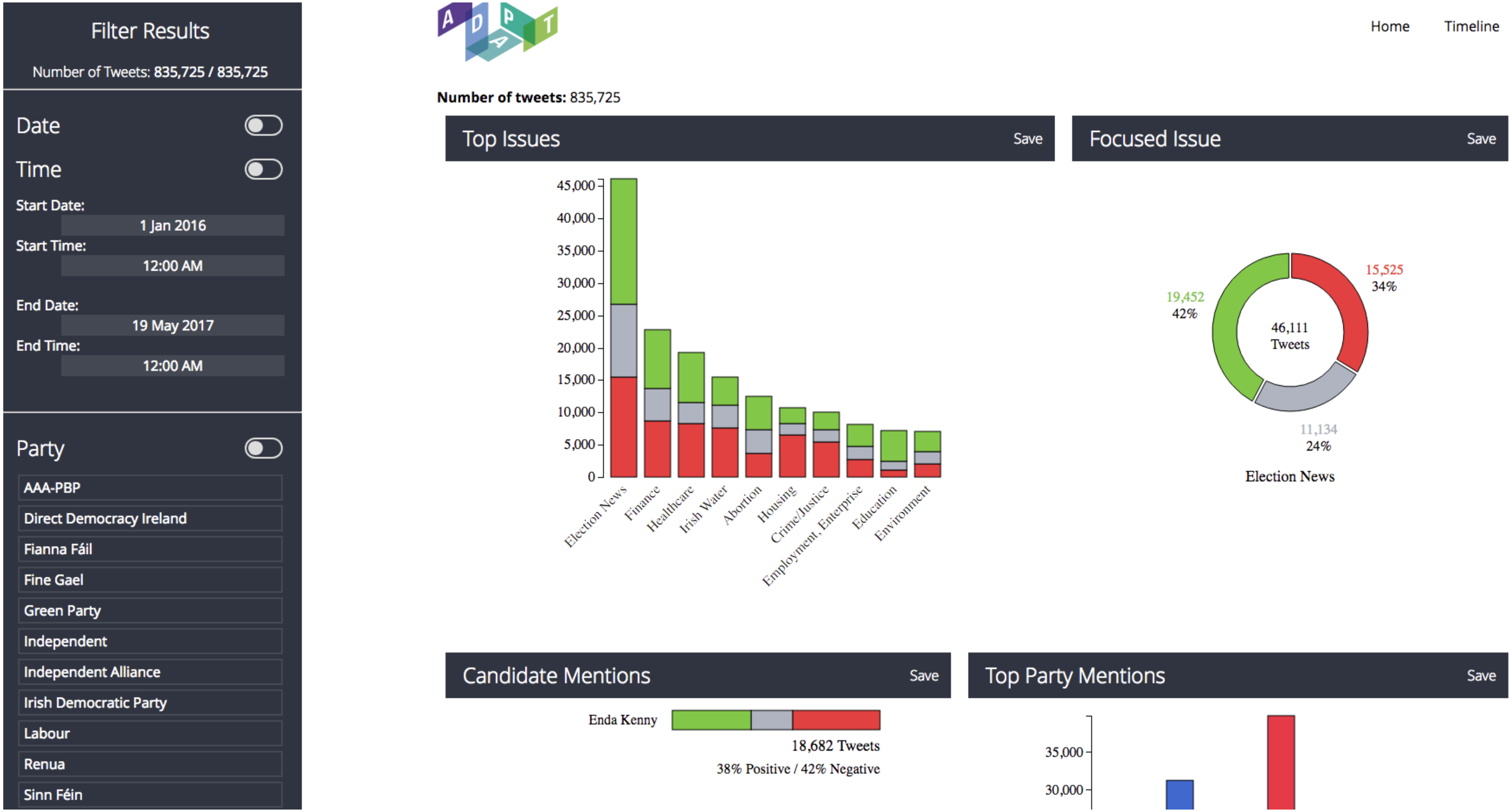
Microsoft → Customer feedback in Japanese

Motivation

Microsoft → Customer feedback in Japanese

Japane
se
data → Translate → Englis
h
data → Sentiment analysis → Sentime
nt
classes

Track Record in UGC



Track Record in UGC

13 languages and 24 language pairs

Brazilator

85,047,110 words in total

Irish	Spanish	Korean
German	Italian	Farsi
French	Portugues e	Greek
Croatian	Japanese	Chinese

↔ English

Sentiment analysis of UGC

- UGC includes blog posts, podcasts, online videos, tweets etc.

- UGC is usually multilingual and of varying quality (sometimes deliberately)

- Sentiment analysis of UGC has many applications

Related work

MT can alter the sentiment (Mohammad et al. (2016))

Google Translate from English to German on 25/05/2017

English: *he is out of the world cup* *negative*

German: *Er ist aus des weltmeisterschaft* *neutral*

Sentiment Analysis of UGC

- Can a sentiment classification approach help improve sentiment preservation in the target language ?

Sentiment Analysis of UGC

- Can a sentiment classification approach help improve sentiment preservation in the target language ?

- Is it useful to select a specific sentiment-MT model to translate the UGC with the same sentiment ?

Data preparation

Corpus development:

- Twitter data set comprising 4,000 English tweets from the FIFA World Cup 2014 and their manual translations into German

Data preparation

Corpus development:

- Twitter data set comprising 4,000 English tweets from the FIFA World Cup 2014 and their manual translations into German

- Informal translations of English tweets into German

e.g. English tweet German tweet

Goaaaal

Tooor

- **Sentiment annotation**

Manually annotated sentiment scores between 0 and 1

- **Sentiment annotation**

Manually annotated sentiment scores between 0 and 1

- **Sentiment classes**

(i) Negative: sentiment score ≤ 0.4

(ii) Neutral: sentiment score ≈ 0.5

(iii) Positive: sentiment score ≥ 0.6

e.g. Tweet Sentiment score

injured Neymar out of World Cup 0.2

Sentiment annotation and classification

- Manual annotation of Twitter data is considered as the “gold-standard”

- Manual annotation of Twitter data is considered as the "gold-standard"
- 50 tweets per sentiment (negative, neutral and positive) are held out for tuning and testing purposes

Data	Train	Development			Test			Total
		#neg	#neu	#pos	#neg	#neu	#pos	
Twitter	3,700	50	50	50	50	50	50	4,000

Data distribution of Twitter data for Training, development and test

Sentiment annotation and classification

- Flickr and News commentary (``News'') data are used as additional resources
- Automatic sentiment analysis tool (Afli et. al. (2017)) is applied to Flickr and News data

Sentiment annotation and classification

- Flickr and News commentary (``News’’) data are used as additional resources
- Automatic sentiment analysis tool (Afli et. al. (2017)) is applied to Flickr and News data

Performance accuracy:

- 2,994 tweets out of 4,000 correctly classified by this tool when compared to the ‘gold standard’ data
- Accuracy = **74.85%**

Sentiment annotation and classification

Data	Sentiment classification	#neg	#neu	#pos	#total
Twitter	manual	919	1,308	1,473	3,700
Flickr	automatic	9,677	11,065	8,258	29,000
News	automatic	111,337	14,306	113,200	238,843

Data distribution after sentiment classification

Experiments

I. **Translation without sentiment classification**

Experiments

I. Translation without sentiment classification

II. Translation with sentiment classification

i. Manual sentiment classification (only Twitter data)

ii. Automatic sentiment classification (Flickr & News data)

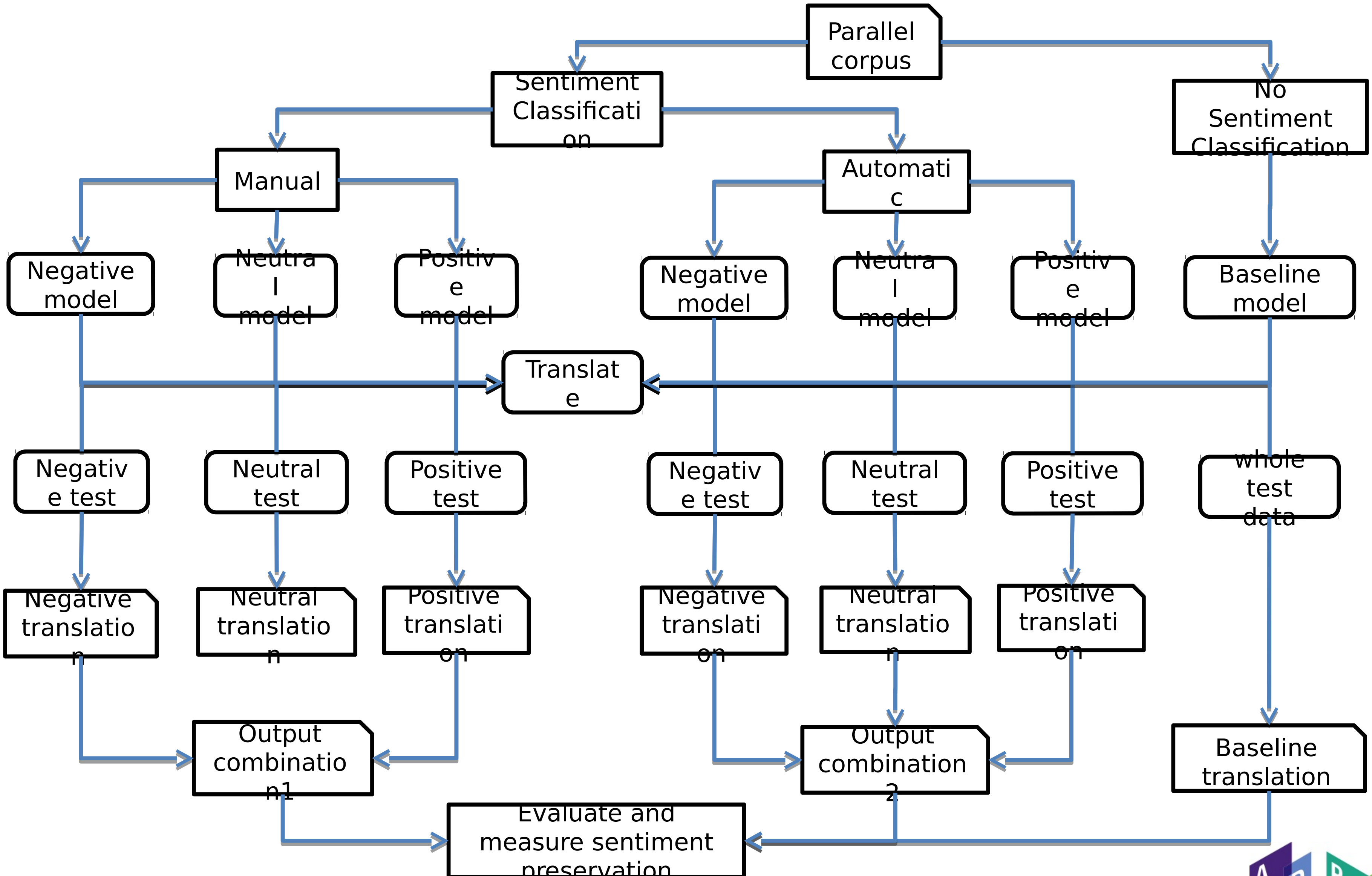
Experiments

I. **Translation without sentiment classification**

II. **Translation with sentiment classification**
   i. Manual sentiment classification (only Twitter data)
   ii. Automatic sentiment classification (Flickr & News data)

III. **Translation by wrong MT engines**
   i. Negative tweets by positive model
   ii. Neutral tweets by negative model
   iii. Positive tweets by neutral model

Sentiment Translation Architecture



Results

Translation model	Data size	Sentiment Classification	BLEU	METEOR	TER	Sentiment Preservation
Twitter	4k	√	48.2	59.4	34.2	72.66%
Twitter (Baseline)		×	50.3	60.9	31.9	66.66%
Twitter + Flickr	33k	√	48.5	59.8	33.9	71.33%
Twitter + Flickr		×	50.7	62.0	31.3	62.66%
Twitter + Flickr + News	272k	√	50.3	62.3	31.0	**75.33**%
Twitter + Flickr + News		×	**52.0**	**63.4**	**30.1**	73.33%
Twitter (Wrong MT engine)	4k	√	46.9	57.9	35.4	47.33%

Experimental evaluation with data concatenation

Results

Translation model	Data size	Sentiment Classification	BLEU	METEOR	TER	Sentiment Preservation
Twitter	4k	√	48.2	59.4	34.2	72.66%
Twitter (Baseline)		×	50.3	60.9	31.9	66.66%
Twitter + Flickr	33k	√	48.5	59.8	33.9	71.33%
Twitter + Flickr		×	50.7	62.0	31.3	62.66%
Twitter + Flickr + News	272k	√	50.3	62.3	31.0	**75.33**%
Twitter + Flickr + News		×	**52.0**	**63.4**	**30.1**	73.33%
Twitter (Wrong MT engine)	4k	√	46.9	57.9	35.4	47.33%

Experimental evaluation with data concatenation

Results

Translation model	Data size	Sentiment Classification	BLEU	METEOR	TER	Sentiment Preservation
Twitter	4k	√	48.2	59.4	34.2	72.66%
Twitter (Baseline)		×	50.3	60.9	31.9	66.66%
Twitter + Flickr	33k	√	48.5	59.8	33.9	71.33%
Twitter + Flickr		×	50.7	62.0	31.3	62.66%
Twitter + Flickr + News	272k	√	50.3	62.3	31.0	**75.33**%
Twitter + Flickr + News		×	**52.0**	**63.4**	**30.1**	73.33%
Twitter (Wrong MT engine)	4k	√	46.9	57.9	35.4	47.33%

Experimental evaluation with data concatenation

Results

Translation model	Data size	Sentiment Classification	BLEU	METEOR	TER	Sentiment Preservation
Twitter	4k	√	48.2	59.4	34.2	72.66%
Twitter (Baseline)		×	50.3	60.9	31.9	66.66%
Twitter + Flickr	33k	√	48.5	59.8	33.9	71.33%
Twitter + Flickr		×	50.7	62.0	31.3	62.66%
Twitter + Flickr + News	272k	√	50.3	62.3	31.0	**75.33%**
Twitter + Flickr + News		×	**52.0**	**63.4**	**30.1**	73.33%
Twitter (Wrong MT engine)	4k	√	46.9	57.9	35.4	47.33%

Experimental evaluation with data concatenation

Results

Translation model	Data size	Sentiment Classification	BLEU	METEOR	TER	Sentiment Preservation
Twitter	4k	√	48.2	59.4	34.2	72.66%
Twitter (Baseline)		×	50.3	60.9	31.9	66.66%
Twitter + Flickr	33k	√	48.5	59.8	33.9	71.33%
Twitter + Flickr		×	50.7	62.0	31.3	62.66%
Twitter + Flickr + News	272k	√	50.3	62.3	31.0	**75.33**%
Twitter + Flickr + News		×	**52.0**	**63.4**	**30.1**	73.33%
Twitter (Wrong MT engine)	4k	√	46.9	57.9	35.4	47.33%

Experimental evaluation with data concatenation

Examples

Example	Reference	Sentiment translation model	Baseline model
1	Howard Web is a terrible ref #WorldCup	Howard Web is a schrecklicher ref #WorldCup	Howard Web is a schrecklicher ref #WorldCup
2	injured Neymar out of World Cup 2014	verletzter Neymar out the WC2014	verletzter Neymar out of World Cup 2014
3	penalty shootouts are too intense !	penalty shoot is to intensiv !	penalties is to intensiv !
4	damn chile is nice !!!! #WorldCup	freeking Chile is good !!! #WorldCup	damn Chile is good !!! #WorldCup
5	a bit boring …	a little boring …	some boring …
6	im with Germany	I stand to Germany's side	I stand to Deutschlands side
7	as getting I, GO CHILE !	completely mache I it GO CHILE !	as getting I, GO CHILE !

Comparison of translations by sentiment translation models and Baseline model

Examples

Example	Reference	Sentiment translation system	Baseline system
1	Bosnia and Herzegovina really got f*** over man	Bosnia and Herzegowina eliminated echt demolished	Bosnia and Herzegovina were a abgezogen
2	when USA lost , but were still moving onto the next round	even if USA today we in the next round	could usa loses the next round
3	Brazil 5 WorldCup championship Argentine 2 WorldCup championship so Ill go with Brazil	Brazil 5 time world champion Argentina 2 time world champion so Im for Brazil	Brazil 5 time world champions Argentina 2 time world champions so for Brazil

Examples where sentiment is altered by the Baseline system

Examples

Example	Reference	Sentiment translation system	Baseline system
1	Bosnia and Herzegovina really got f*** over man	Bosnia and Herzegowina eliminated echt demolished	Bosnia and Herzegovina were a abgezogen
2	when USA lost , but were still moving onto the next round	even if USA today we in the next round	could usa loses the next round
3	Brazil 5 WorldCup championship Argentina 2 WorldCup championship so Ill go with Brazil	Brazil 5 time world champion Argentina 2 time world champion so Im for Brazil	Brazil 5 time world champions Argentina 2 time world champions so for Brazil

Examples where sentiment is altered by the Baseline system

Examples

Example	Reference	Right MT engine	Wrong MT engine
1	little break on the #WorldCup for an amazing #Wimbledon final!	small Pause from the #WorldCup for a amazing #Wimbledon final!	kleine Pause of the #WorldCup for a erstaunliches #Wimbledon final!
2	yes !!!!!	yes !!!!!	so !!!!!
3	a bit boring ...	a little boring ...	some was ...

Comparison between sentiment polarities using the right and wrong MT engine

Discussion

- MT scores are better when no sentiment classification is used
- Sentiment classification approach performs better than for systems where it is switched off

Discussion

Translation model	Sentiment Classification	BLEU	Sentiment Preservation
Twitter	√	48.2	72.66% (+6%)
Twitter (Baseline)	×	50.3 (+2.1)	66.66%
Twitter + Flickr	√	48.5	71.33% **(+8.67%)**
Twitter + Flickr	×	50.7 (+2.2)	62.66%
Twitter + Flickr + News	√	50.3	**75.33**% (+2%)
Twitter + Flickr + News	×	**52.0** (+1.7)	73.33%

MT quality VS *sentiment preservation*

Discussion

- In most cases, the Baseline system produces better outputs in terms of BLEU score
- In some cases, interestingly, sentiment classification approach produces better outputs

Discussion

- In most cases, the Baseline system produces better outputs in terms of BLEU score
- In some cases, interestingly, sentiment classification approach produces better outputs
- Using specific sentiment-MT model to translate a text with the same sentiment is better in both ways

Translation model	Sentiment Classification	BLEU	Sentiment Preservation
Twitter (Right MT engine)	√	48.2 (+1.3)	72.66% (+25%)
Twitter (Wrong MT engine)	√	46.9	47.33%

MT quality vs. *sentiment preservation*

Sentiment translation system using nearest classes:

1. Combine the negative- and neutral-sentimented parallel twitter data

The nearest sentiment class approach

Sentiment translation system using nearest classes:

1. Combine the negative- and neutral-sentimented parallel twitter data
2. Combine the neutral- and positive-sentimented parallel twitter data

The nearest sentiment class approach

Sentiment translation system using nearest classes:

1. Combine the negative- and neutral-sentimented parallel twitter data
2. Combine the neutral- and positive-sentimented parallel twitter data
3. Build the “negative-neutral” translation model using the data in **Step 1**

The nearest sentiment class approach

Sentiment translation system using nearest classes:

1. Combine the negative- and neutral-sentimented parallel twitter data
2. Combine the neutral- and positive-sentimented parallel twitter data
3. Build the “negative-neutral” translation model using the data in **Step 1**
4. Build the “positive-neutral” translation model using the data in **Step 2**

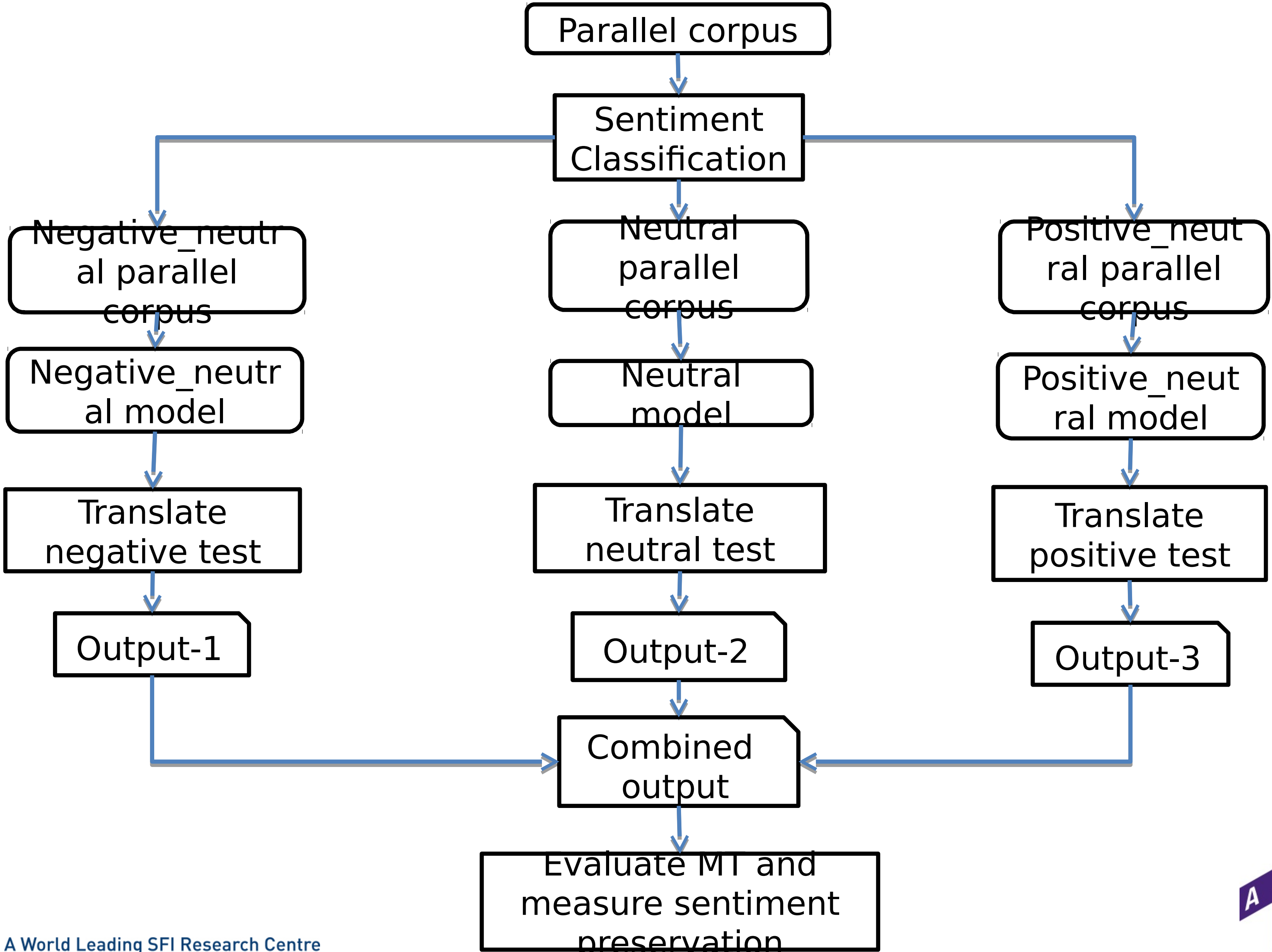
Experiments

- Building a single translation system using the whole twitter data
- Building the sentiment-specific translation systems
- Building the nearest-sentiment translation systems

Data distribution:

Experiments	Train	Development			Test			Total
		#neg	#neu	#pos	#neg	#neu	#pos	
Exp 1	3,700	50	50	50	50	50	50	4,000
Exp 2	3,400	100	100	100	100	100	100	4,000

Architecture: nearest sentiment class



Results

Exp1 (150 test data)	BLEU	Meteor	TER	Sentiment Preservation
Twitter (Baseline)	**50.3**	**60.9**	**31.9**	66.66%
Twitter_SentClass	48.2 (-2.1)	59.4 (-1.5)	34.2 (+2.3)	**72.66%** (+6)
Twitter_NearSent	49.0 (-1.3)	60.1 (-0.8)	34.0 (+2.1)	66.66% (+0)

Results

Exp1 (150 test data)	BLEU	Meteor	TER	Sentiment Preservation
Twitter (Baseline)	**50.3**	**60.9**	**31.9**	66.66%
Twitter_SentClass	48.2 (-2.1)	59.4 (-1.5)	34.2 (+2.3)	**72.66%** (+6)
Exp2 (300 test data)	**BLEU**	**Meteor**	**TER**	**Sentiment Preservation**
Twitter (Baseline)	**51.3**	**62.5**	**31.0**	52.33%
Twitter_SentClass	47.3 (-4)	59.1 (-3.4)	35.2 (+3.8)	**60.33%** (+8)
Twitter_NearSent	48.3 (-3)	59.6 (-2.9)	34.4 (+3.4)	60.0% (+7.67)

Results

Exp1 (150 test data)	BLEU	Meteor	TER	Sentiment Preservation
Twitter (Baseline)	**50.3**	**60.9**	**31.9**	66.66%
Twitter_SentClass	48.2 (-2.1)	59.4 (-1.5)	34.2 (+2.3)	**72.66%** (+6)
Exp2 (300 test data)	**BLEU**	**Meteor**	**TER**	**Sentiment Preservation**
Twitter (Baseline)	**51.3**	**62.5**	**31.0**	52.33%
Twitter_SentClass	47.3 (-4)	59.1 (-3.4)	35.2 (+3.8)	**60.33%** (+8)
Twitter_NearSent	48.3 (-3)	59.6 (-2.9)	34.4 (+3.4)	60.0% (+7.67)

***Note that scores are different for Exp1 and Exp2 as they are tested on different data distributions

Discussion

- A small of amount of test data (i.e., 150 tweet pairs) is not (really) sufficient to confirm our hypothesis

- Comparatively larger test data (i.e., 300 tweet pairs) confirms the utility of our approach

- Our approach is able to reduce the gap between translation quality and sentiment preservation

Conclusions

- Despite a small deterioration in translation quality, our approach significantly improves sentiment preservation
- It is essential to carefully select the proper MT engine conveying the same sentiment polarity as that of the UGC
- The nearest sentiment class-inclusion approach helps improve the balance between MT quality and sentiment preservation

Future work

- Apply to other language pairs and also other forms of UGC such as customer feedback, blogs etc.

- Further refine the sentiment classes (strong positive, strong negative etc.) in order to build more specific translation models

Thank you