A World Leading SFI Research Centre

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

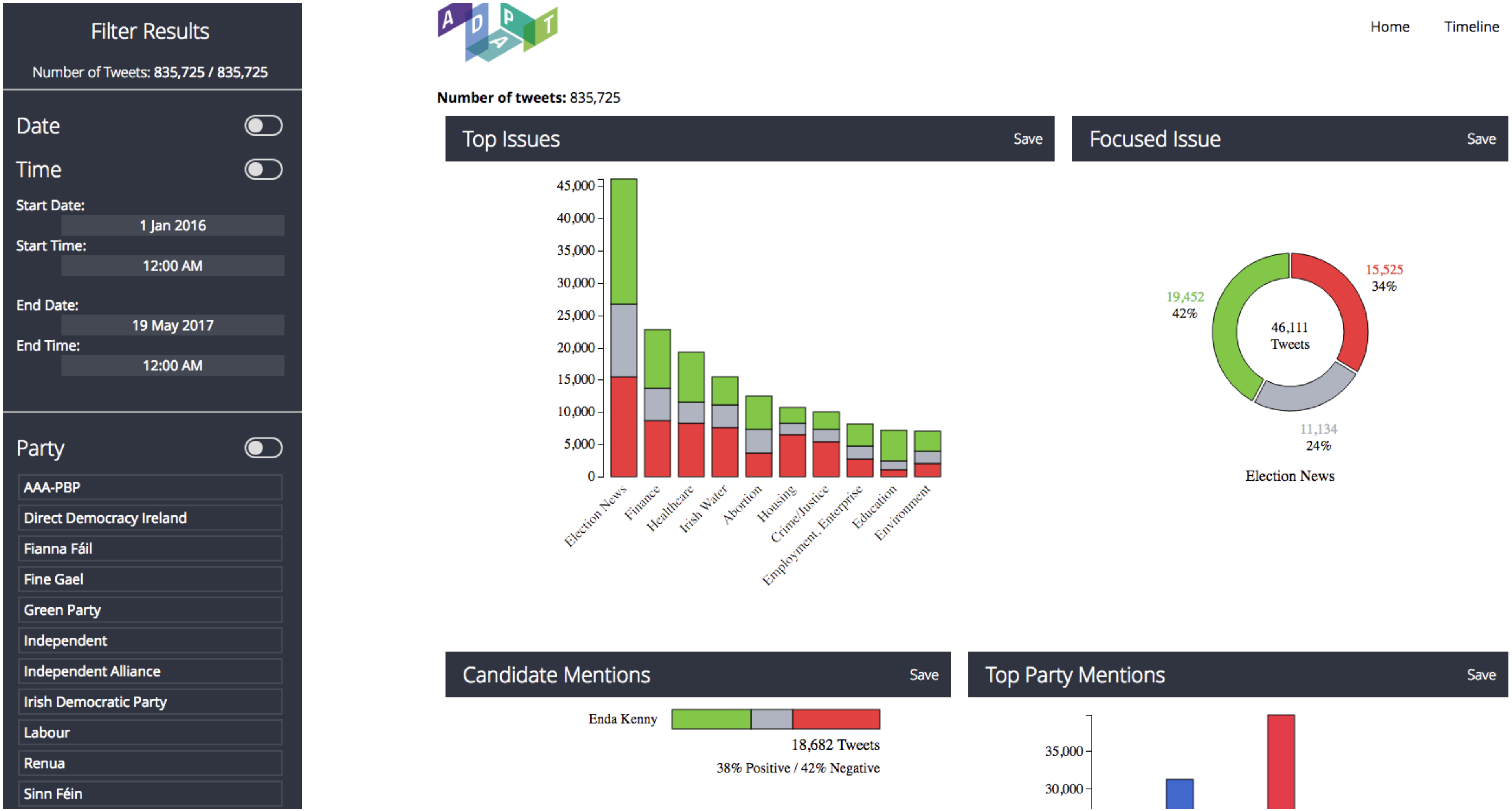
Microsoft → Customer feedback in Japanese

Motivation

Microsoft → Customer feedback in Japanese

Japanese data → Translate → English data → Sentiment analysis → Sentiment classes

Track Record in UGC



Track Record in UGC

13 languages and 24 language pairs

Brazilator

85,047,110 tweets in total

Irish	Spanish	Korean
German	Italian	Farsi
French	Portuguese	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

Sentiment analysis of UGC

Crosslingual sentiment analysis(CLSA):

- The task of predicting the polarity of the opinion of a text in a language using a classifier trained on the corpus of another language (Balamurli et al. (2012))

Sentiment analysis of UGC

Crosslingual sentiment analysis(CLSA):

- The task of predicting the polarity of the opinion of a text in a language using a classifier trained on the corpus of another language (Balamurli et al. (2012))

MT-based CLSA:

- MT is utilized to leverage its capability, existing SA resources available in English to classify sentiment in other languages (Mihalcea et al. (2012))

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-sentimented 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 *Toooor*

Sentiment annotation and classification

- **Sentiment annotation**

  Manually annotated sentiment scores between 0 and 1

Sentiment annotation and classification

- **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”

Sentiment annotation and classification

- 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**

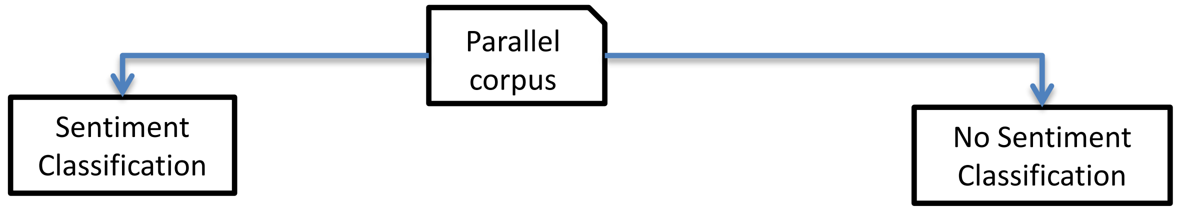
II. **Translation with sentiment classification**
   i. Manual sentiment classification (only Twitter data)
   ii. Automatic sentiment classification (Flickr & News data)

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

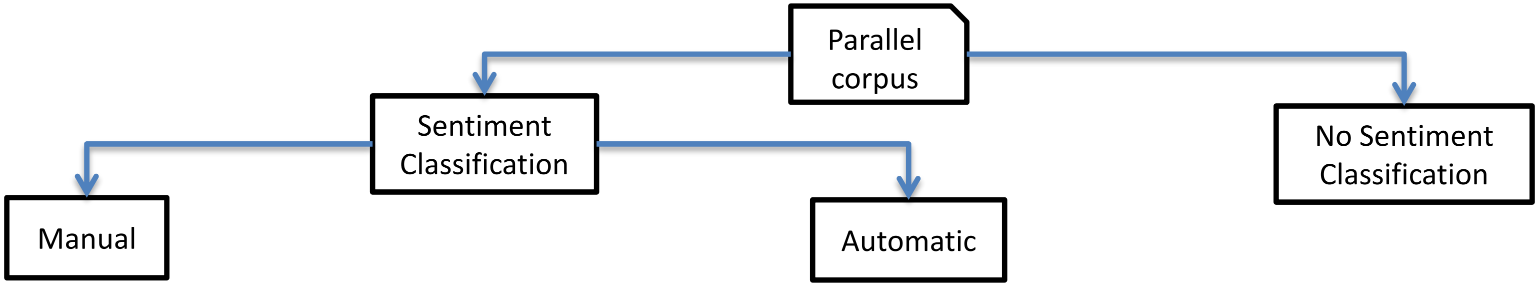
Sentiment Translation Architecture

Parallel corpus

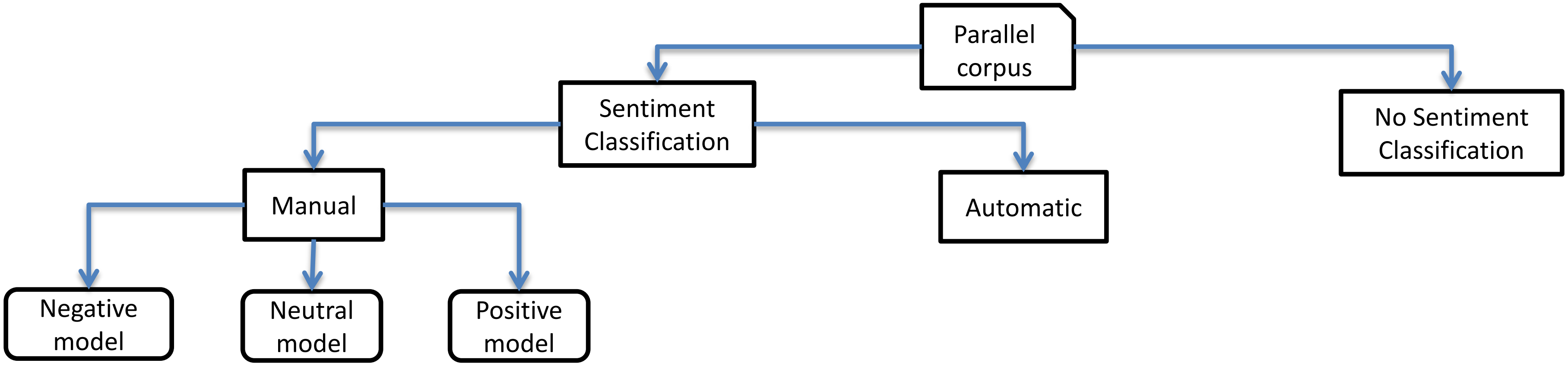
Sentiment Translation Architecture



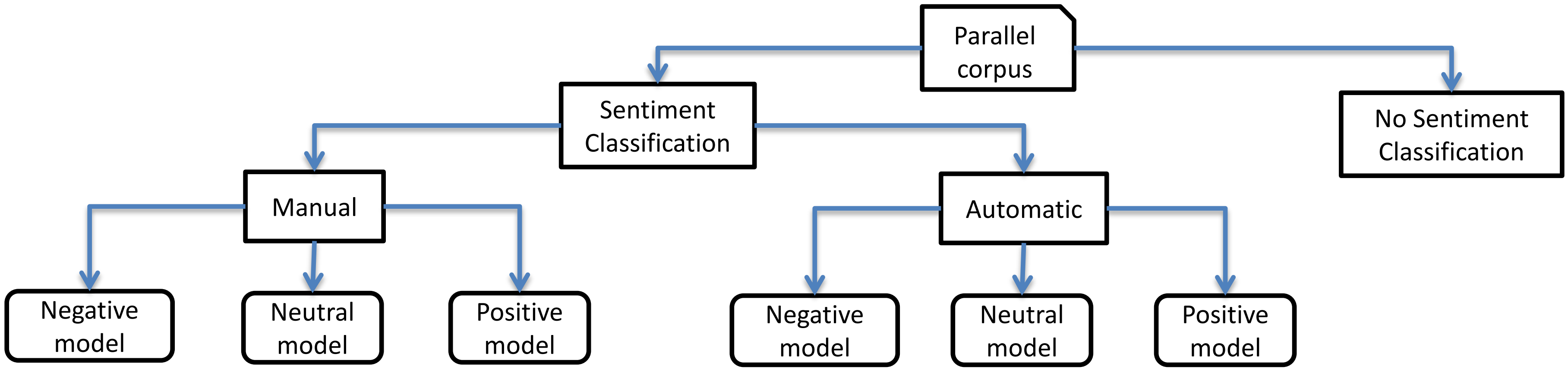
Sentiment Translation Architecture



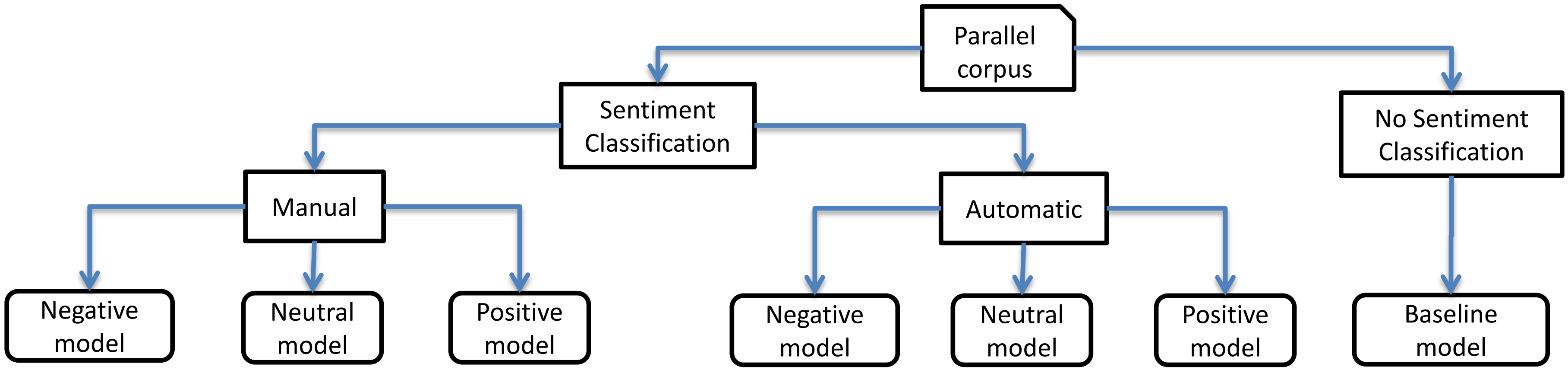
Sentiment Translation Architecture



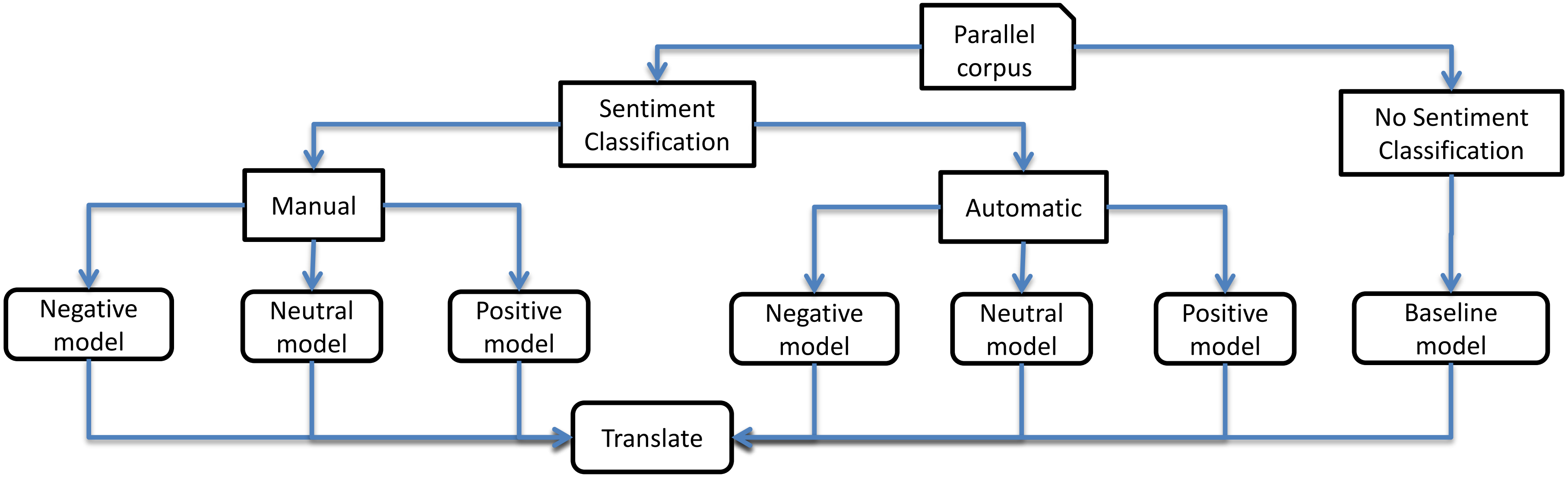
Sentiment Translation Architecture



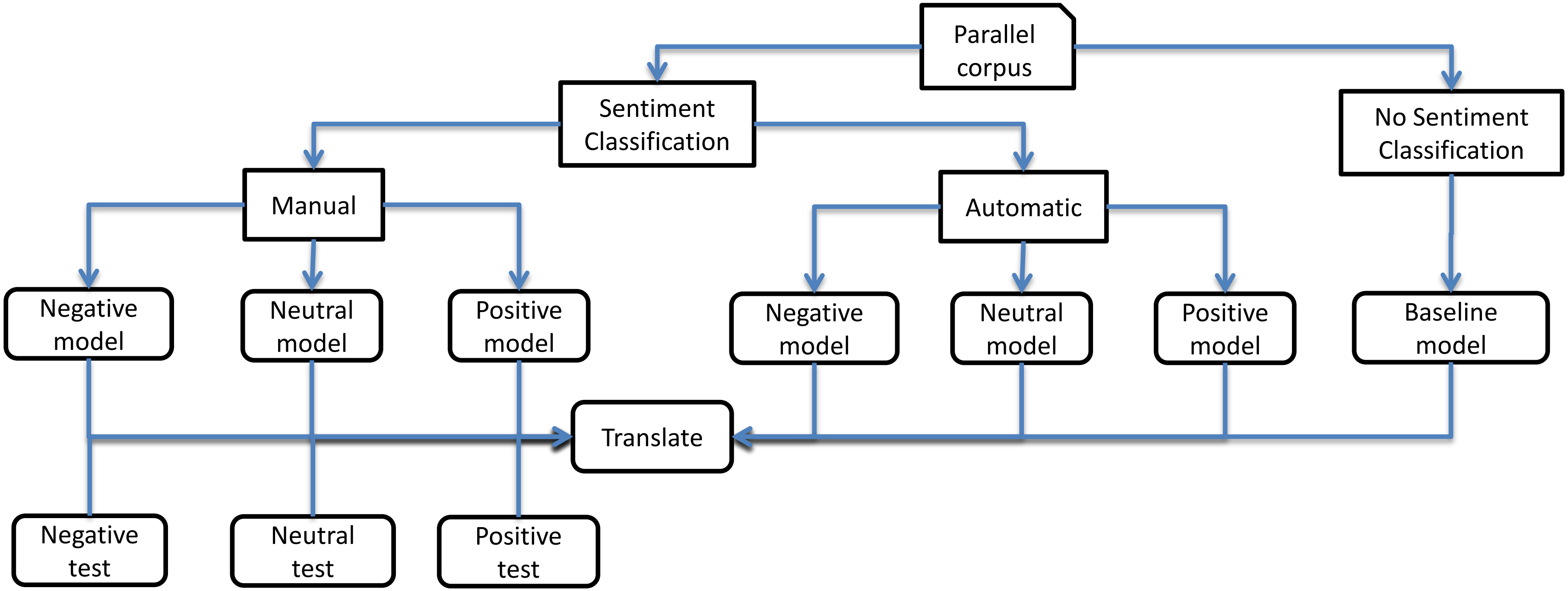
Sentiment Translation Architecture



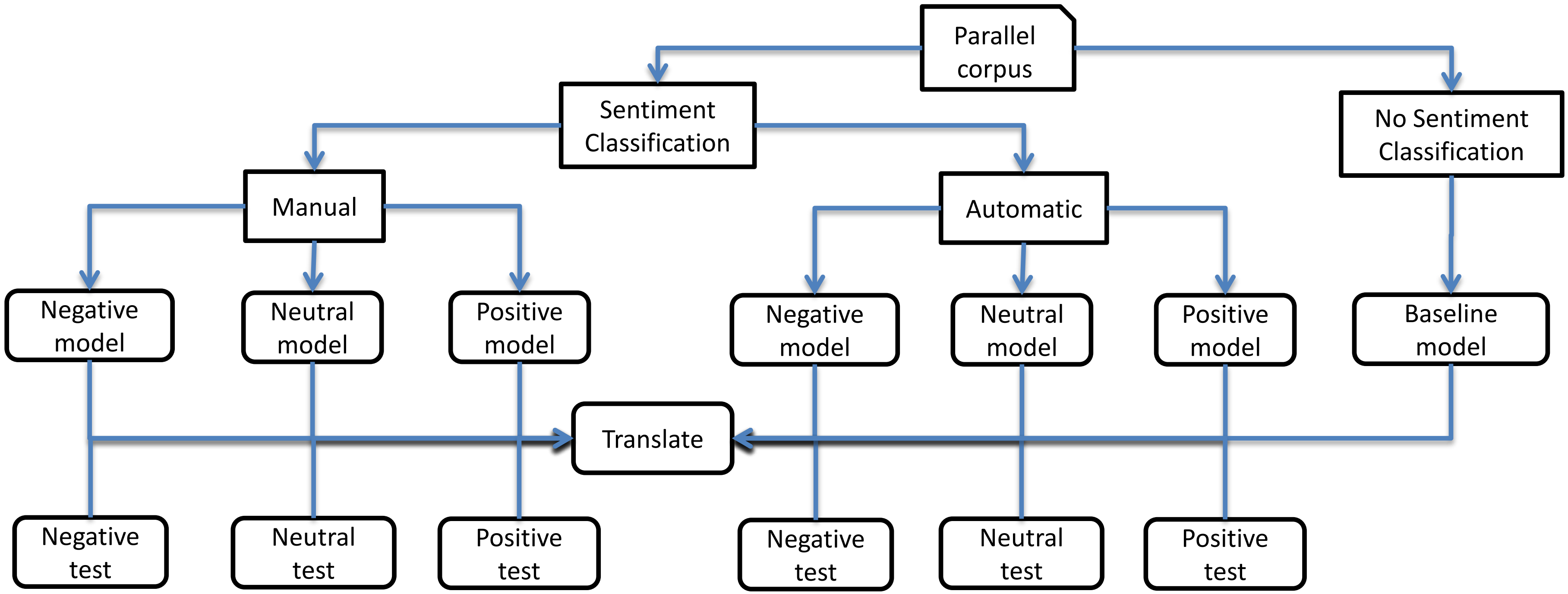
Sentiment Translation Architecture



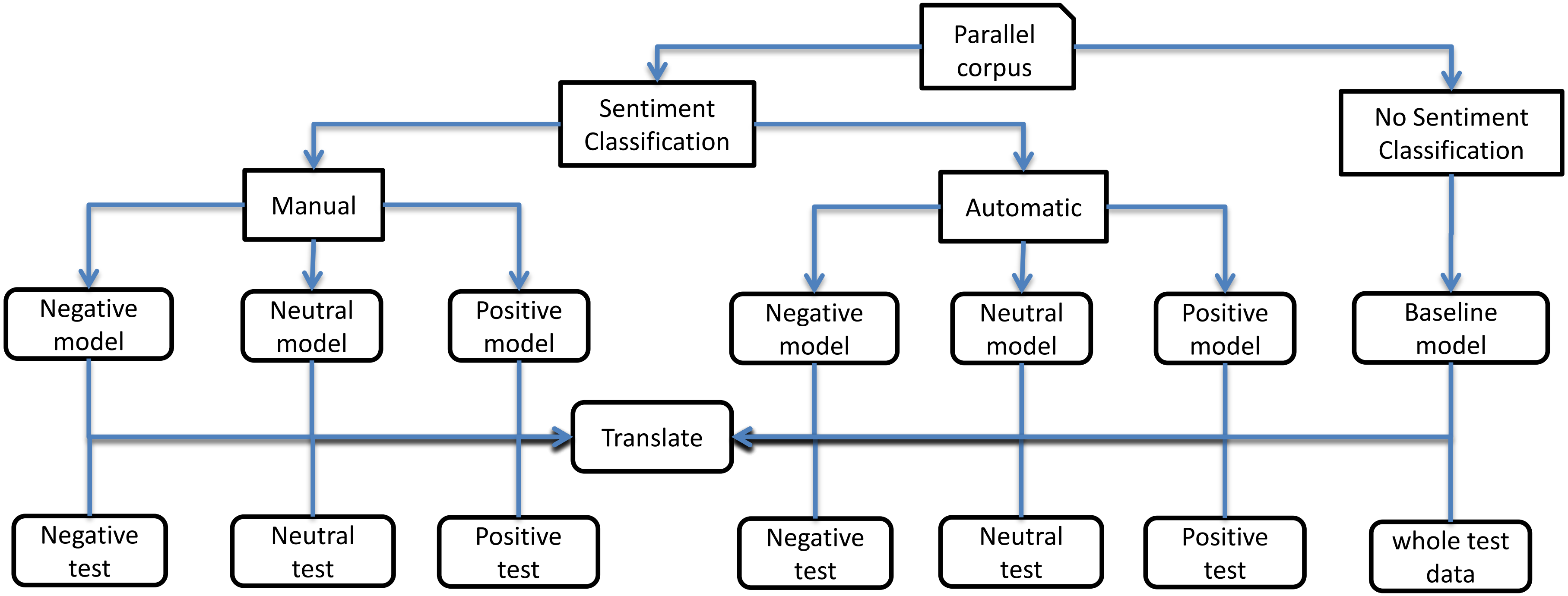
Sentiment Translation Architecture



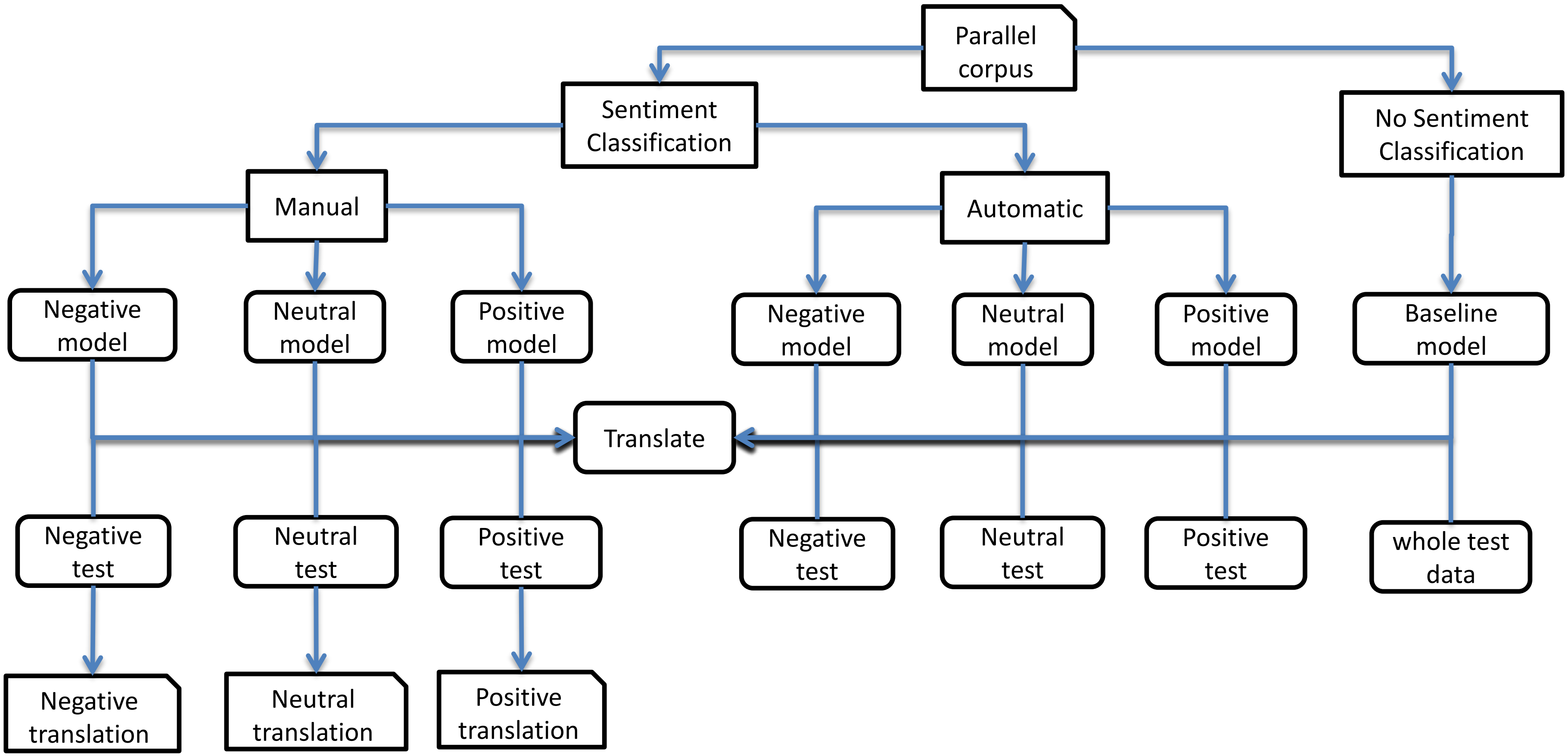
Sentiment Translation Architecture



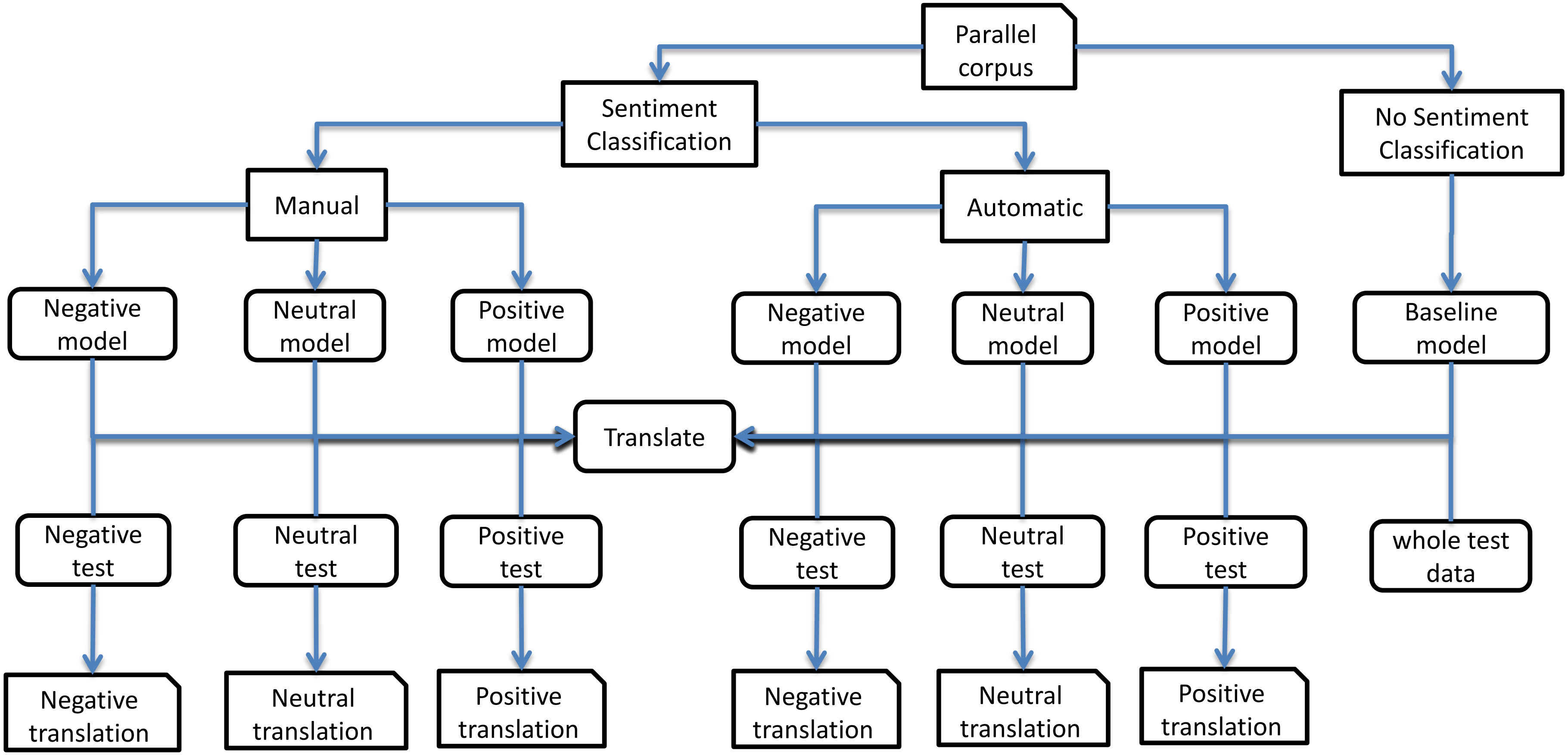
Sentiment Translation Architecture



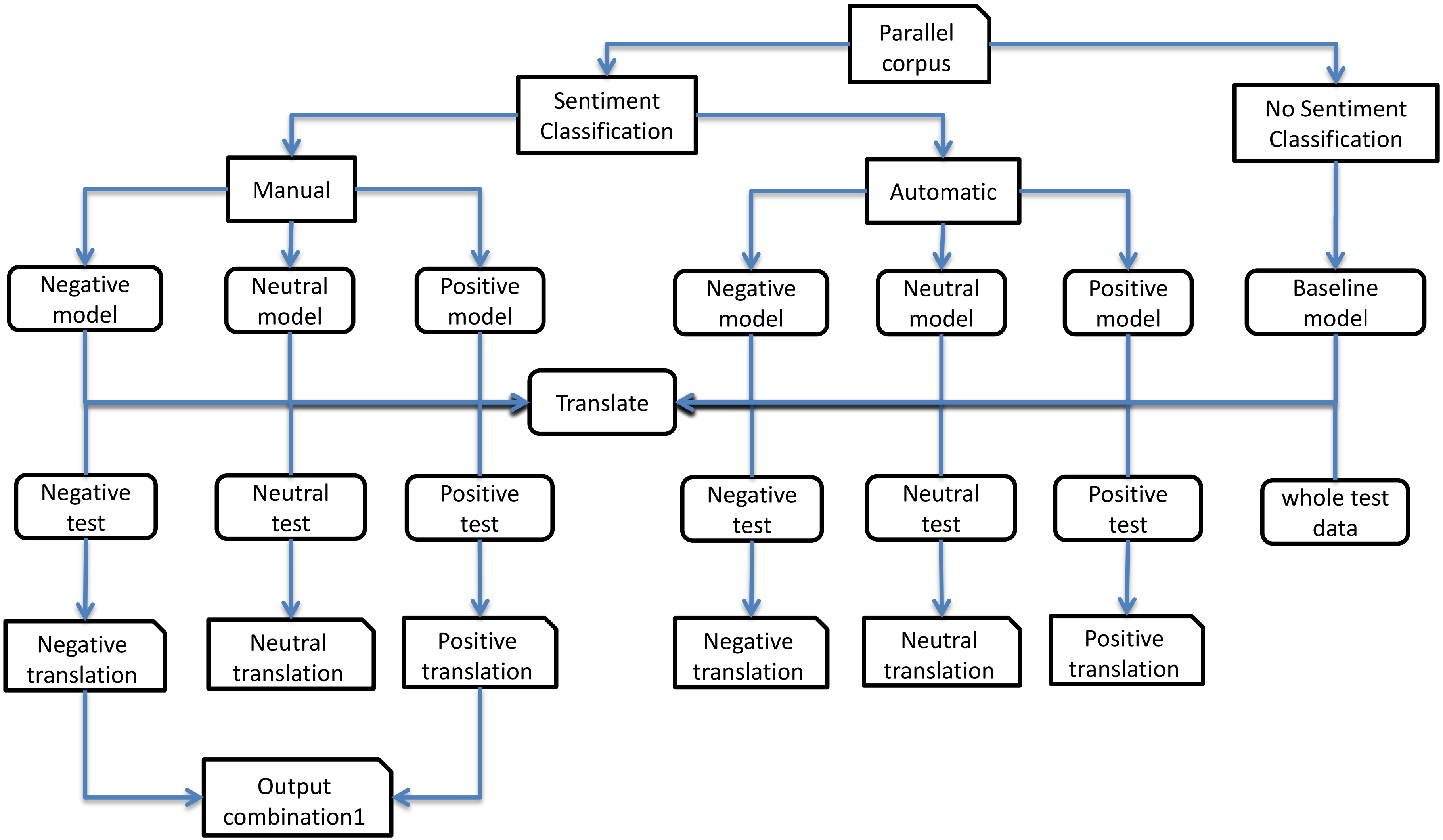
Sentiment Translation Architecture



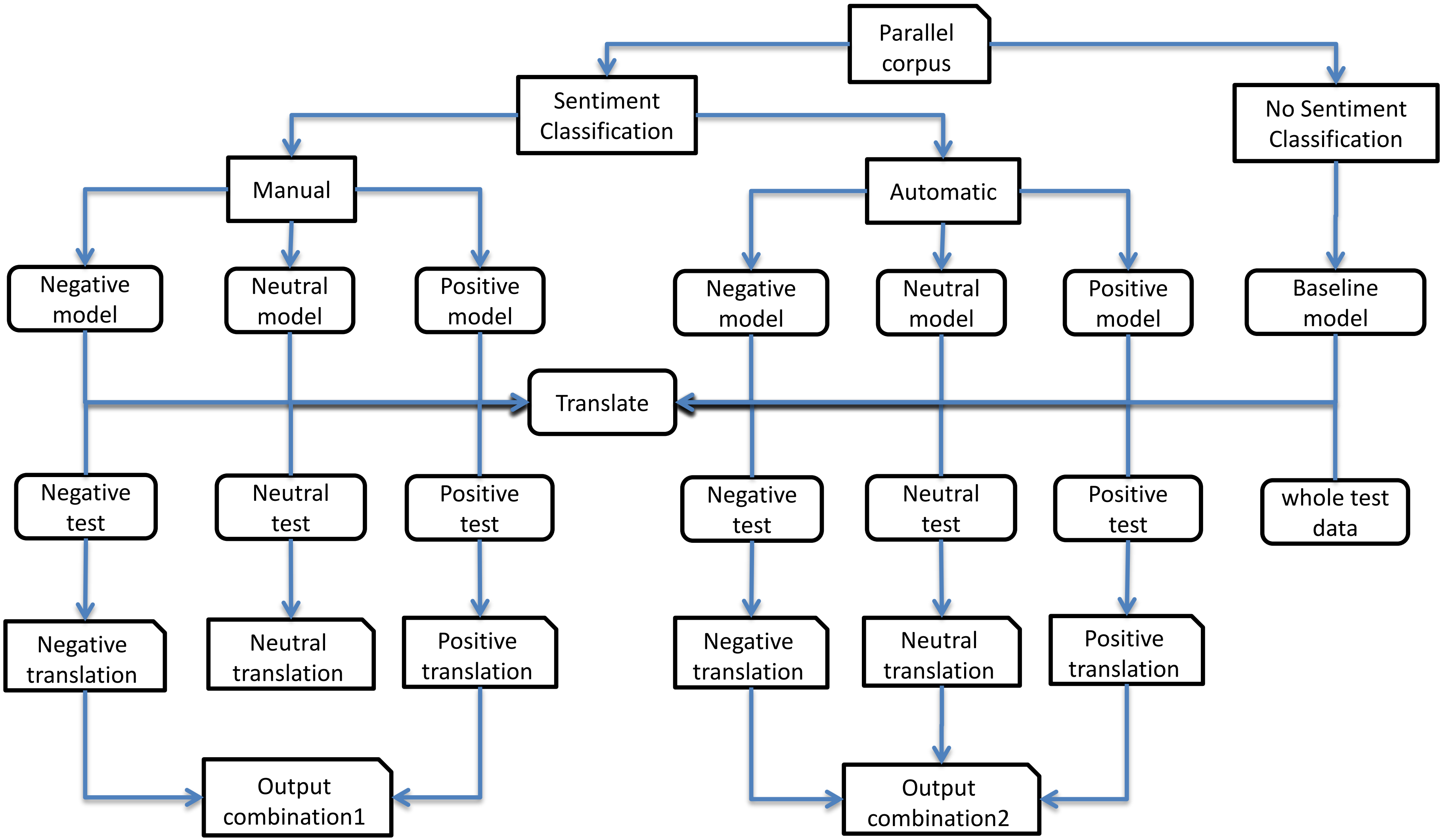
Sentiment Translation Architecture



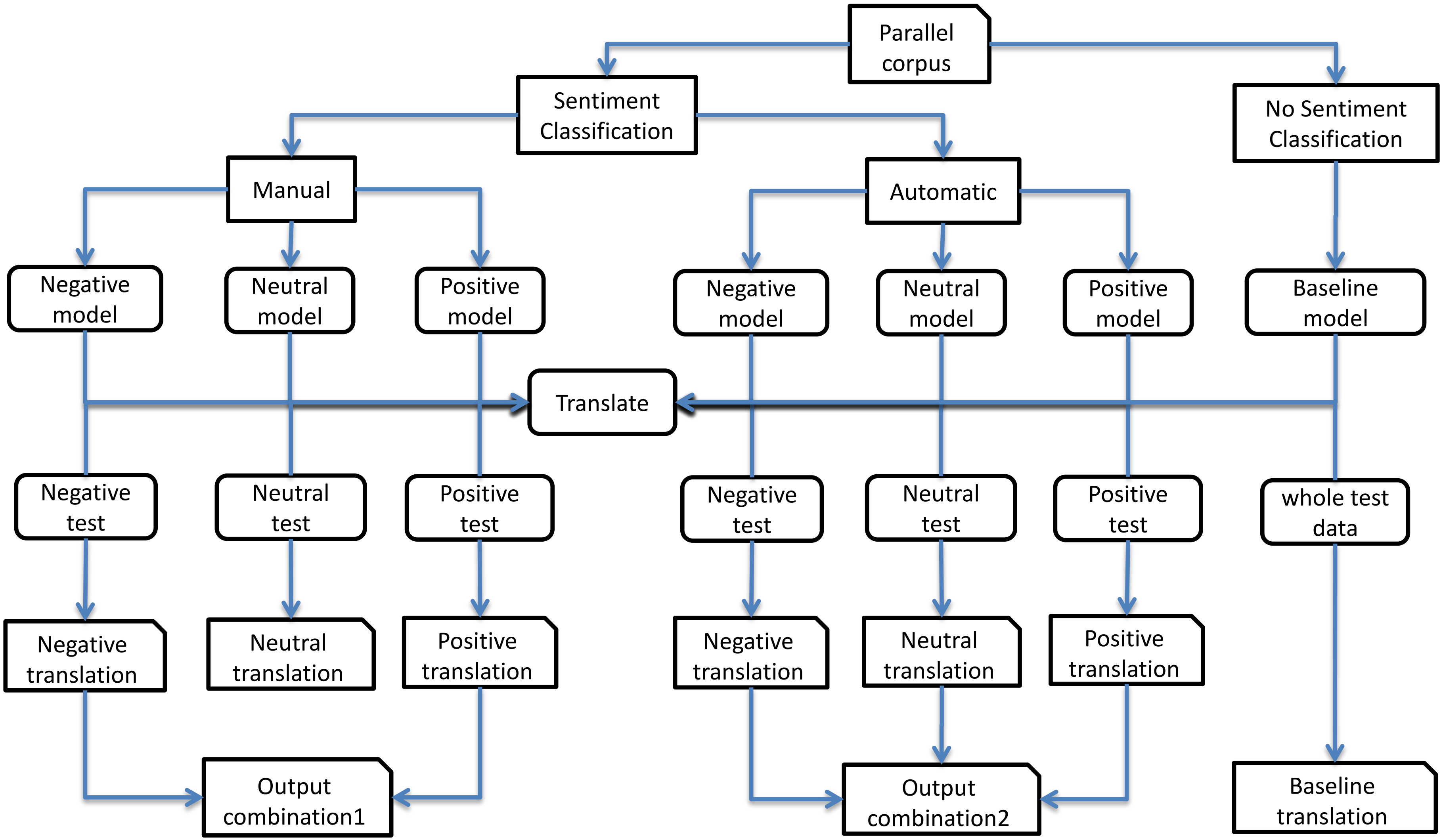
Sentiment Translation Architecture



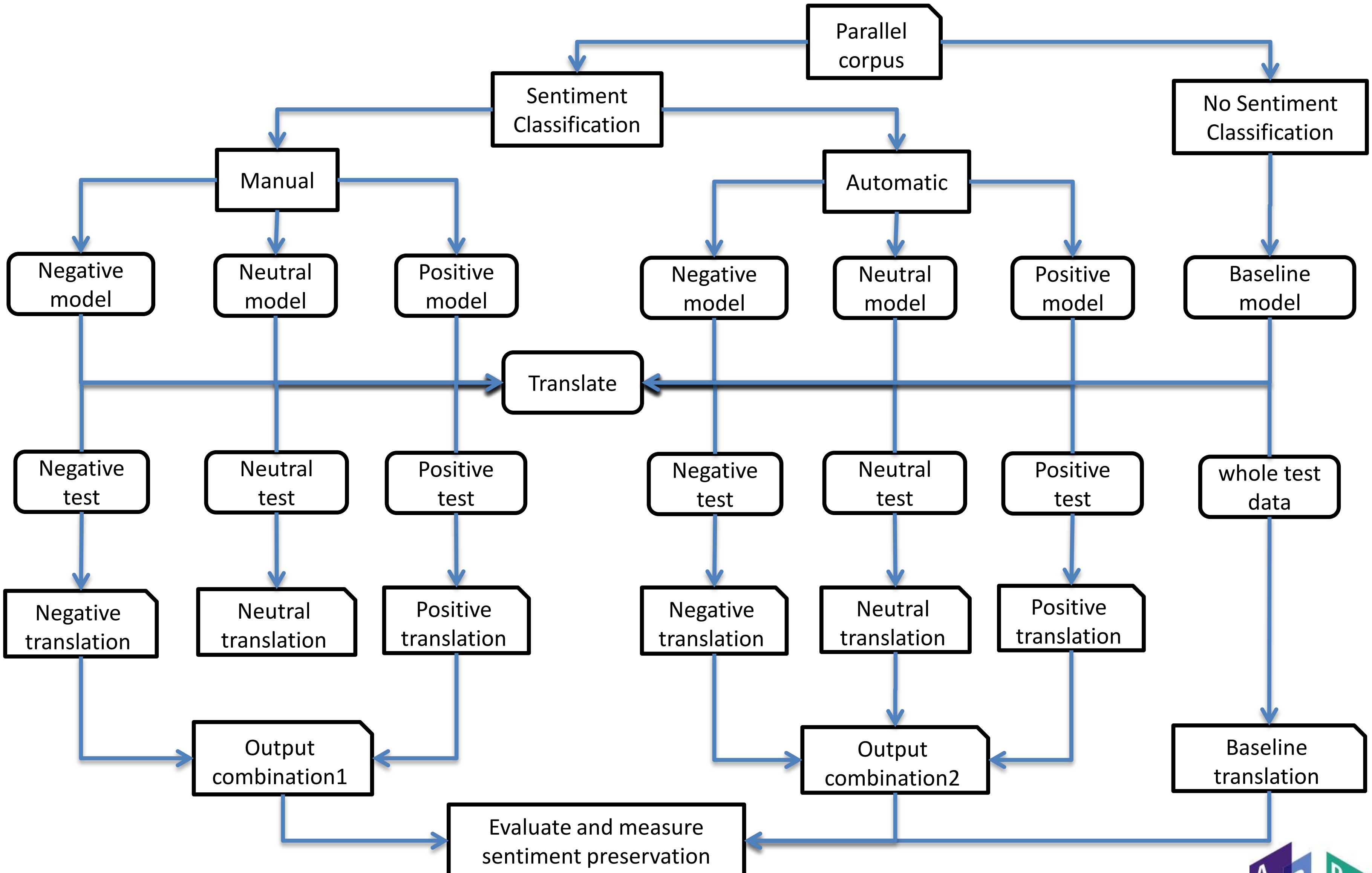
Sentiment Translation Architecture



Sentiment Translation Architecture



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 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 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 Deutschlands side	I stand to Germany's 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 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 the 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-sentimented 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*

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

Future work

- To 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

pintu.lohar@adaptcentre.ie