Features for Modelling Characteristics of Conversations

Gunnar Eriksson   Jussi Karlgren

2012-09-20
Gavagai AB

Stockholm, Sweden

What information to use?

What information to use?

- what content in conversations?

What information to use?

- what content in conversations?
- what behaviour within conversations?

What information to use?

- what content in conversations?
- what behaviour within conversations?
- what types of conversations?

Feature types

Feature types

Conversation content

> words and short phrases

Feature types

Conversation content
  words and short phrases

Conversation behaviour
  turn-taking – the flow of interaction

Feature types

Conversation content
> words and short phrases

Conversation behaviour
> turn-taking – the flow of interaction

Conversation type
> number of participants and conversation length

Author profile

Train a maximum entropy classifier with all features from all conversations the author took part in. Some exceptions:

- do not use features with occurrence with only one author.
- remove feature frequency information.

Lexical features

Lexical features

- normalisation and tokenisation
- tokens and token bigrams
- bag of tokens + bigrams for every participant

Lexical features

- normalisation and tokenisation
- tokens and token bigrams
- bag of tokens + bigrams for every participant
- generalize mentions of chat nick-names of other conversation partakers to **OtherName**.
- generalize mentions of nick-name of the "speaker" to **SelfName**.

More than one source to the content of a conversation!

SLEX The things you utter in the conversation.

OLEX The things others utter in the conversation.

More than one source to the content of a conversation!

SLEX The things you utter in the conversation.

OLEX The things others utter in the conversation.

Add the two sets of features to every author profile.

In two ways: Concatenation (SLEX + OLEX) or union (CLEX).

Combination of lexical sources

Features types used	Precision	Recall	F (β=1)
submitted	0.84	0.89	0.87
SLEX	0.21	0.84	0.33
OLEX	0.27	0.80	0.35
OLEX+SLEX	0.43	0.85	0.57
CLEX	0.48	0.72	0.58
CLEX + CONV features	0.56	0.55	0.55

Conversation type, CTYPE

Conversation type, CTYPE

Monologue

0 − 1 participant

Dialogue

2 participants

Group

> 2 participants

Conversation length, LTYPE

Conversation length, LTYPE

Feature	# of utterances
fail	0 – 1
handshake	2 – 7
prelude	8 – 25
brief	26 – 50
discourse	51 – 100
ldiscourse	101 – 160
vldiscourse	161 –

Adding ctype to ltype

Example

$prelude_{mono}$, $prelude_{dia}$, or $prelude_{group}$
 $brief_{mono}$, $brief_{dia}$, or $brief_{group}$

...

Turn-taking behaviour, TTAKE

The flow of utterances within a conversation is modelled by turn-taking trigrams from each participant's perspective.

Turn-taking behaviour, TTAKE

The flow of utterances within a conversation is modelled by turn-taking trigrams from each participant's perspective.

Example

	utterance 1	u2	u3	u4	u5	...
participants	A	A	B	A	B	...

Example

A: Self–Self–Other, Self–Other–Self, Other–Self–Other, ...
B: Other–Other–Self, Other–Self–Other, Self–Other–Self, ...

Adding ctype to ttake

Adding ctype to ttake

Example

A: Self–Self–Other$_g$, Self–Other$_g$–Self, Other$_g$–Self–Other$_g$, ...
B: Other$_g$–Other$_g$–Self, Other$_g$–Self–Other$_g$, Self–Other$_g$–Self, ...

Combining it all

Features types used	Precision	Recall	F (β=1)
S-OLEX+CTYPE+LTYPE+TTAKE	0.84	0.89	0.87
CTYPE+LTYPE+TTAKE	0.68	0.04	0.06
S-OLEX+CTYPE+LTYPE	0.90	0.90	0.90
S-OLEX+LTYPE+TTAKE	0.95	0.93	0.94
S-OLEX+CTYPE+TTAKE	**0.97**	**0.97**	**0.97**

Take home

- it takes two to tango – use that!
- and even simplistic situational features helps!

Normalisation and tokenisation

All lower-case

1. let all sequences of black characters delimited by white-space constitute a token.

2. let all initial and final sequences of punctuation characters be a token of its own.

3. add extra tokens from (some) tokens with internal black-space: URL $\rightarrow$ URL + URL parts.

Identifying flagrant utterances

1. use the sexual predator classifier.

2. classify all predator utterances in training material using only SLEX features.

3. rank the utterances by the probability for the utterance "being a sexual predator".

4. decide a flagrancy threshold.

1. classify all utterances in test material as above.

2. pick all utterances above threshold.