Alignment and Emotionality in Dialogue

Sara Veldhoen
10545298
sara.veldhoen@student.uva.nl

Abstract

Alignment is observed in dialogue on several levels. This research reviews the effect of emotionality of the interlocutors on their tendency to align to each other. A corpus of forum scrapes is used, containing mutli-party written dialogue. Alignment is measured on a lexical and syntactic level. Emotionality is approximated using annotations of sarcasm, niceness, and appealing to fact or feeling, amongst others. No clear correlation is found between these emotionality measures and linguistic alignment, but some of the results require further investigation.

1 Introduction

People tend to converge to the same behaviour in interaction, which is referred to as *alignment*. In dialogue, interlocutors align on different linguistic levels such as phonology, lexical choice, and syntax. From a communicative point of view, alignment on lower levels is driven by the aim to reach semantic alignment, i.e. to maximize mutual understanding. We align to the linguistic behaviour of our interlocutor to make sure they are able to understand what we say. Therefore, we need to maintain some kind of *model* of them.

When people are very emotional, they seem to be less capable of adopting another perspective. This would cause the model of their interlocutor to be a poorer representation of them. My hypothesis is therefore that people align less in conversation when they are very emotional. I will investigate this hypothesis with a focus on negative emotions such as anger, irritation, and grumpiness.

Other perspectives on alignment do not require an explicit model of our interlocutors. Alignment can also be viewed as an automatic process that emerges from our cognitive architecture, for instance stemming from neurological priming. In that case, it may or may not be influenced by our emotional state.

For my experiments, I use the Internet Argument Corpus (AIC) as introduced in (Walker et al., 2012). The corpus contains debates on politics from Internet forums. The corpus was designed for argumentation research. It has some emotionality annotated, I give a detailed description of the dataset in section 3

2 Related work

The assumption underlying my hypothesis is that people would be less capable of adopting another perspective when emotional. Indeed, research has shown that emotionality constrains cognitive resources for several tasks (Storbeck, 2012). It is argued that emotion serves as a cue to prioritize cognitive processes. More specifically, negative emotion negatively affects a relational processing style, and performance on verbal tasks goes down.

2.1 Alignment

A lot of research has focused on alignment in discourse. In (Reitter and Moore, 2010) the relationship between alignment and task success is studied. The HCRC Map Task corpus is used for this, which contains spoken interaction on a clear defined cooperative task.

Alignment is measured on both lexical and syntactic level. First, each utterance is parsed into a constituent structure. The repetition of context-free production rules is counted (for syntactic alignment) as well as repetition of words (lexical alignment).

A distinction is made in (Reitter and Moore, 2010) between short-term and long-term alignment effects. The former, called priming, is a strong effect with a quick decay: a low plateau is reached within seconds after the stimulus. The latter can last up to minutes or even days. Short-term priming is measured by counting repetition of a

prime in a sliding window over 15 seconds. Long-term priming is measured by cutting the dialogue in two: the first half is treated as priming period, and the second half as target. A prior for rule and lexical repetitions is estimated on a control case, which combines unrelated dialogue halves.

Another research is aimed at the correlation between alignment of linguistic style and power or status (Danescu-Niculescu-Mizil et al., 2012). In this case, alignment is quantified by counting occurrence of certain categories of function words. The occurrence function words, that have little semantic meaning, reflects linguistic style rather than content of the message. The baseline probability of using such words is estimated on conversation between the same interlocutors, and only the direct influence in the next reply is counted as coordination.

2.2 Emotionality in discourse

In (Justo et al., 2014), an attempt is made to classify emotions like sarcasm and nastiness in forum posts. These properties of utterances were annotated in part of the AIC corpus, thus enabling supervised learning. Nastiness can be detected rather easily using surface patterns such as abusive terms. Detecting sarcasm is less trivial and requires a combination of statistical cues, linguistic and semantic information. The semantic information is obtained using an existing dictionary of words related to 64 categories, among which are some emotions.

3 Data

The AIC corpus contains forum scrapes, rather than spoken dialogue. This has a number of consequences:

- There are generally more than two interlocutors involved in a conversation. This is very much related to my hypothesis, which is about modeling the interlocutor. You could say that in a forum, people do not model a specific interlocutor but instead a more general audience. This is by itself an interesting phenomenon for the communicative stance to review.

- Obviously, it is not possible to look at alignment on a phonological or phonetic level in written dialogue.

Topic	Items
abortion	1299
climate change	44
communism vs. capitalism	34
death penalty	51
evolution	1635
existence of God	308
gay marriage	610
gun control	1092
health care	88
marijuana legalization	34
Total	5195

Table 1: Number of post pairs extracted for each topic

- There is no syntactic annotation in the corpus. But I expect utterances in written dialogue to be sentential more often. Therefore, it is possible to do syntactic parsing, in order to look at syntactic alignment.

- One property of so-called planned (written) conversation as opposed to spontaneous speech, is that the language used is generally richer. This may result in lower alignment overall.

Unfortunately, there is no annotation of poster identity in the data, so I cannot consider people's personal tendency to align.

As a side mark, the forum scrapes in the AIC are from the website `fourforums.com`. The audience is not a random sample from the population, but consists of people who are attracted to forum discussions. It is good to note that this may be reflected in their attitude and emotional behavior.

3.1 Annotations

Part of the AIC is annotated with Mechanical Turk for topic, agreement, and some other interesting properties. The annotations were performed in two distinct tasks, by approximately 6 Turkers for each item. An overview of the annotated properties is presented in table 2.

The annotations were performed on posts in response to some quoted post. In the current experiments, the quoted post is taken as the stimulus that can incite alignment.

I have extracted 5195 post pairs in nine topics, see table 1. In a substantial part (574) of those post pairs, the task 2 annotations are not available.

Task 1		
agreement	Does the respondent agree or disagree with the prior post?	scalar
fact-feeling	Is the respondent attempting to make a fact based argument or appealing to feelings and emotions?	scalar
attack	Is the respondent being supportive/respectful or are they attacking/insulting in their writing?	scalar
sarcasm	Is the respondent using sarcasm?	binary
nicenasty	Is the respondent attempting to be nice or is their attitude fairly nasty?	scalar
Task 2		
agree-disagree	Does the respondent agree or disagree with the previous post?	binary
negotiate-attack	Does the respondent seem to have an argument of their own OR is the respondent simply attacking the original poster's argument?	scalar
defeater-undercutter	Is the argument of the respondent targeted at the entirety of the original poster's argument OR is the argument of the respondent targeted at a more specific idea within the post?	scalar
questioning-asserting	Is the respondent questioning the original poster OR is the respondent asserting their own ideas?	scalar
personal-audience	Is the respondent's arguments intended more to be interacting directly with the original poster OR with a wider audience?	scalar

Table 2: Mechanical-Turk annotations in the IAC

4 Experimental set-up

4.1 Alignment measures

Each annotated or *focus* post has a post that it responds to: the stimulus. I compute the baseline probability of the focus post occurring at all, and also the probability of it occurring after the stimulus. The relative increase of post probability from the baseline is a measure of the alignment, see equation 1.

$$Alignment = \frac{P(post|stimulus)}{P(post|baseline)} \quad (1)$$

If this value is 1, no alignment is observed at all. The lower this value, the more alignment there is. I compute these probabilities on two levels: lexical and syntactic.

Note that all equations in this section are presented as regular probabilities, but for efficiency and numerical stability the actual computations are made in log-space.

Syntactic alignment For the syntactic dimension, all sentences are parsed with the Stanford Parser in the `nltk` package. The resulting trees are used to construct a probabilistic context free grammar. This is a grammar that associates each context-free production (or *rule*) with a certain probability. I will simply use the maximum likelihood estimate, as in equation 2.

$$P(A \rightarrow B) = \frac{Count(A \rightarrow B)}{\sum_{B'} Count(A \rightarrow B')} \quad (2)$$

The probability of a post given a grammar model is obtained by multiplying the probability of the parses of the sentences (equation 3). The probability of a parse tree is the product of its rule applications (equation 4). Since we are comparing scores for which the sentences and associated parses of the focus post are fixed, there is no need to compensate for the length of the post or the trees.

$$P(p|GM) = \prod_{s \in p} P(parse(s)|GM) \quad (3)$$

$$\text{where } P(t|GM) = \prod_{r \in t} P(r|GM) \quad (4)$$

Lexical alignment The lexical dimension concerns word-choice. A unigram language model is created, by counting occurrences of words and taking the maximum likelihood estimate. Note that I do not use any stemming: a process in which

inflectional suffixes are removed from words to obtain a more general lemma. This could be added to the model to reduce data sparsity effects, but it may also conceal useful information. Note that stemming does not have a big effect on function words, that are more representative of linguistic style anyway, according to (Reitter and Moore, 2010).

The probability of a post given a lexical model is obtained by taking the product of its lexical items: equation 5. Again, since the content the post is fixed, there is no need to compensate for the length.

$$P(p|LM) = \prod_{s \in p} \prod_{w \in s} P(w|LM) \qquad (5)$$

Back-off The local models are trained on restricted datasets, which causes the estimates for many events to be equal to zero. In such cases, the estimate from a back-off model is used. For the syntactic estimation, the back-off model is trained on all post pairs in the corpus. For the lexical estimation, a post-specific back-off model is trained on all posts belonging to the same topic. This is done in order to somewhat factor out the content of the posts. For instance, in a conversation about abortion, the probability of the word 'baby' occurring is simply higher than in general. I do not mean to measure this as alignment.

Note that neither back-off model will have a zero estimate for the items in the posts under review, as these posts are included in the training material.

Reliability In general, a relative frequency estimate is less stable with few datapoints. This makes the local (post-based) models less reliable. For instance, if there is only one sentence in the stimulus post, there is only one rule with S as a root, so the estimate for this rule will be 1.0, whereas it would be 0.5 in case there are two sentences.

This effect is present but less dramatic for the lexical models, since their is no fine-grained classification of words, as opposed to syntactic rules that are estimated per root/ left hand side. Still, the relative frequency estimate is sensitive to data sparsity. Apart from the local models, also the baseline lexical models for topics with few posts suffer from this effect.

4.2 Alignment and emotionality

Of course, there is no unambiguous classification of emotionality of the poster in the corpus. The annotations may however serve as a proxy for (negative) emotions. In particular the 'sarcasm','nice/nasty', and 'fact-feeling' annotations (all from task 1) seem very useful. I also add 'attack', 'agreement' from task 1, and'personal-audience', 'agree-disagree', 'questioning-asserting', and 'defeater-undercutter' from task 2 to the set of relevant features to look at. I use the average score of all available annotations for an item (from different annotators).

Furthermore, I use capitalization of entire words (of length more than one) and the occurrence of emoticons in posts as other proxies for emotionality. The former is a conventional substitute for yelling, as illustrated in the following snippet from `106720.post` about abortion:

> "well i was downplaying your argument because really, pregnancy is not that demeaning to a womans body. ITS NATURAL! OBESITY ISNT! Which issues? The health issues?"

Emoticons are explicitly meant to express emotionality in forums. The distribution of them is not very high in this dataset, which is why I consider all emoticons equal instead of using a fine-grained count per emoticon or class of emoticons.

I estimate correlation between the alignment measures explained in section 4.1, and the measures of emotionality detailed above.

Correlation is estimated as Pearson's ρ, which yields values ranging from -1 (exact negative correlation) via 0 (no correlation) to 1 (exact positive correlation). The relation is negligible roughly between -0.19 and 0.19.

With Pearson's ρ, only linear correlation is measured. Although I do not necessarily expect a strictly linear relation between alignment and emotionality, there is no a priori reason to assume a higher-order relationship.

5 Results and analysis

In general, there is almost always some alignment effect observed. Lexical alignment is 0.82 on average, syntactic alignment 0.77. Recall that 1.0 means no alignment, and smaller values correspond to increasing alignment. There are some

cases (1 for lexical, 11 for syntactic alignment) in which the stimulus-based model actually results in a *worse* explanation than the baseline, i.e. the alignment is bigger than 1. Manual inspection did not reveal any particular reason for this.

Nicely, there is a high correlation between lexical and syntactic alignment: Pearson's $\rho = 0.801$. This is what we would expect: according to the theory, alignment on lower levels is a prerequisite for higher-level alignment.

However, both alignment measures also correlate quite strongly with the length of the quoted post and focus post. The correlation with quoted post length stems from the reliability issues mentioned in section 4.1: a shorter stimulus post results in extremer estimates in the local model and therefore a higher alignment effect is observed.

The correlation with focus post length is probably also related to this. The longer the focus post is, the more rule applications occur. But the distribution of applicable local-model rules vs. the need to back-off should be independent of focus post length. I cannot really explain this correlation.

For the remainder of this writing, I will refer to 'alignment' as the average of measured lexical and syntactic alignment.

Table 3 displays the correlation of annotated features and the alignment. For the left column, all data points were included. Because of the weaker reliability of topics with fewer posts mentioned in section 4.1, I present the same results computed only for the posts in the abortion topic (that has many posts) in the right column. These figures follow the same trend.

Apart from the correlation with post length already mentioned, the correlation is at negligible levels. There appears not to be a relationship between the measures for emotionality chosen, and the alignment.

In figures 1-3 scatter plots are presented for some emotion-features vs. alignment. Only the abortion-posts are included in the data for these plots.

The nature of the data distribution in the figures is different because the emotionality measures are not on the same scale: sarcasm was originally a binary classification, whereas fact-feeling was annotated on a scale from -5 to 5. The emoticon values are counted occurrences divided by the length of the post.

	all		abortion	
lengthQP	0.375		0.393	
lengthFOCUS	0.347		0.417	
capitals/length	0.023	*	0.003	*
emoticons/length	-0.106		-0.095	
sarcasm	-0.095		-0.106	
nicenasty	0.019	*	0.033	*
fact-feeling	0.175		0.139	
attack	0.002	*	0.004	*
personal-audience	0.080		0.069	*
agreement	-0.145		-0.139	
agree-disagree	-0.105		-0.109	
questioning-asserting	0.117		0.089	
defeater-undercutter	-0.073		-0.090	

Table 3: Correlation (Pearson's ρ) of several features with alignment for entire dataset (left column) and only posts about abortion (right column). Starred values (*) are not significant for $p = 0.1$, the rest is.

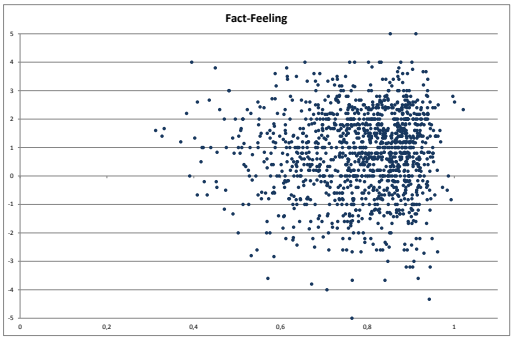

Figure 1: Alignment vs. fact-feeling, for the abortion posts

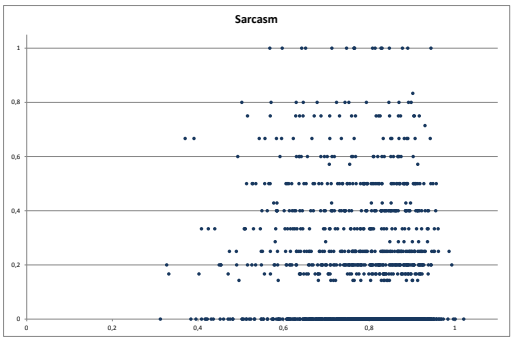

Figure 2: Alignment vs. sarcasm, for the abortion posts

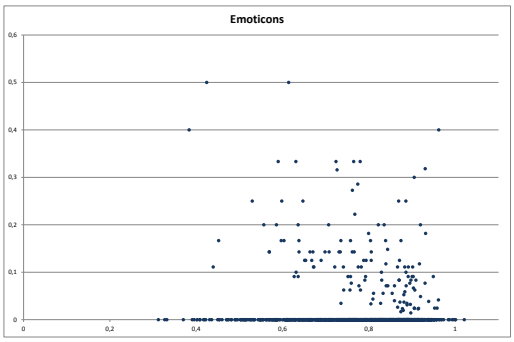

Figure 3: Alignment vs. emoticons, for the abortion posts

Indeed, no correlation is visually discernible in figures 1 and 2. Neither arises a reason to try and fit higher-order relations to the data.

Notably, in figure 3 we see that most data points are on the horizontal axis, i.e. no emoticons are used at all. Unfortunately, there is no way to distinguish post without emoticons because the poster would a priori not use them, and the usage that is actually related to the emotional state of the poster. However, if we focus only on those posts where there is at least one emoticon used, we actually observe a negative correlation between usage of emoticons and alignment. The more emoticons are used, the more emotional the poster is supposed to be, and the less he seems to align. The correlation coefficient is indeed $-0,472$ (significant, strong correlation) if we exclude the zero values.

A closer inspection into the usage of capitals along the same lines reveals the same phenomenon: correlation goes down from 0.003 (hardly any correlation) to -0.285 (significant, weak but noticeable correlation).

These findings do suggest that people align less when emotional. However, no conclusions can be drawn because these effects are distorted from excluding *all* zero emoticon posts from the analysis. Still, it is an interesting observation that deserves closer inspection which is left for other research.

6 Conclusions

There is no strong correlation between any of the emotionality proxies and the alignment as measured on a lexical and syntax level. The results on capitalization and usage of emoticons do however point in the direction of a relationship, but no reliable conclusion can be drawn from this. Future research is needed to see whether these are really related to alignment.

These findings cannot substantiate the hypothesis that alignment goes down when people are emotional. Neither can this hypothesis be rejected because of the results: the measures used may not be a good proxy for emotionality, or emotionality may not occur enough in the dataset.

Future research on this topic could also focus on the semantic features mentioned in section 2.2.

Some doubts remain on the correctness of the alignment measure, as the relative frequency estimate is unstable for small datasets. It is not clear whether this has had a great impact on the present analysis.

References

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708. ACM.

Raquel Justo, Thomas Corcoran, Stephanie M Lukin, Marilyn Walker, and M Inés Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*.

David Reitter and Johanna D Moore. 2010. Predicting success in dialogue.

Justin Storbeck. 2012. Performance costs when emotion tunes inappropriate cognitive abilities: Implications for mental resources and behavior. *Journal of Experimental Psychology: General*, 141(3):411.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).