Measuring Lexico-Semantic Alignment in Debates with Contextualized Word Representations

Aina Garí Soler, Matthieu Labeau, Chloé Clavel
LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
{aina.garisoler,matthieu.labeau,chloe.clavel}@telecom-paris.fr

Abstract

Dialog participants sometimes align their linguistic styles, e.g., they use the same words and syntactic constructions as their interlocutors. We propose to investigate the notion of lexico-semantic alignment: to what extent do speakers convey the same meaning when they use the same words? We design measures of lexico-semantic alignment relying on contextualized word representations. We show that they reflect interesting semantic differences between the two sides of a debate and that they can assist in the task of debate's winner prediction.

1 Introduction

It is well known that dialog participants often tend to imitate each other. This phenomenon, known as alignment or entrainment, can be of a linguistic nature (lexical (Brennan and Clark, 1996), syntactic (Branigan et al., 2000), prosodic (Street Jr, 1984)...) and it has also been observed in non-linguistic behavior such as posture (Shockley et al., 2003) or visual attention (Richardson et al., 2008). For example, throughout a conversation, speakers may reuse the lexical items used by their partners (Nenkova et al., 2008), and they tend to use the same referring expressions to refer to the same entities (Brennan and Clark, 1996). This mechanism is said to facilitate language production and comprehension in the interaction (Pickering and Garrod, 2004); and lexical and syntactic repetition have been found to correlate with task success in task-oriented dialog (Reitter and Moore, 2007).

One kind of alignment that is less often addressed in the literature is conceptual alignment (Stolk et al., 2016). This refers to the extent to which two dialog participants "mean the same things when using the same words" (Schober, 2005). The fact that words have pre-established senses does not guarantee conceptual alignment, as speakers may have slightly different mental representations of words (e.g., different associations, connotations, or a different level of detail), use them differently, or propose novel usages.

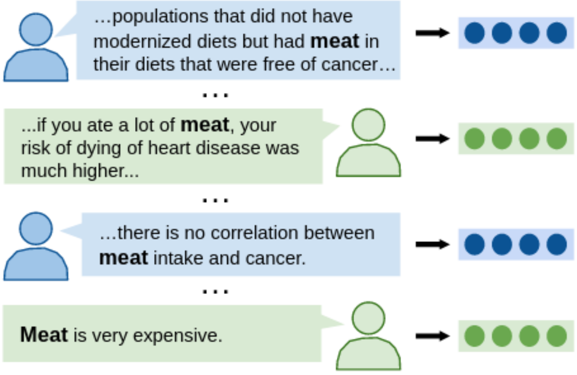
...populations that did not have modernized diets but had **meat** in their diets that were free of cancer...
...
...if you ate a lot of **meat**, your risk of dying of heart disease was much higher...
...
...there is no correlation between **meat** intake and cancer.
...
Meat is very expensive.

Figure 1: We identify words that are used by both sides in a debate (here, *meat*) and extract contextualized representations from all their instances, which are then compared through our alignment measures. Example from the IQ2 dataset (Zhang et al., 2016).

While it has been found that alignment at one level enhances alignment at other levels (Cleland and Pickering, 2003; Pickering and Garrod, 2004), lexical (or surface form) alignment may actually mask conceptual misalignment, which, if undetected, can lead to serious misunderstandings (Schober, 2005).[1] Nevertheless, conceptual (mis)alignment remains understudied, mainly because it is hard to detect.

In this paper, we target a more restricted notion of conceptual alignment: we seek to quantify the divergence or convergence of word meaning that is inferrable from textual information alone, i.e., from the way the same words are used by two speakers throughout a dialog (see Figure 1). We do not intend to capture conceptual misalignments that are made apparent only through non-linguistic

[1] People responding to the same survey twice in the space of a week were twice more likely to change their answers (22% vs 11%) if interviewers had the right to provide clarifications the second time around (Conrad and Schober, 2000). The change in responses indicates that the additional explanations helped uncover and correct an initial conceptual misalignment.

Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023), pages 50–63
July 14, 2023 ©2023 Association for Computational Linguistics

behavior or which involve external referents (e.g., someone performing the wrong action after misunderstanding a command). We refer to this notion as **lexico-semantic alignment**.

We propose, for the first time, a methodology and a set of metrics to explore and quantify lexico-semantic alignment in its definition presented above. Our metrics rely on contextualized word representations, which have been found to reflect different aspects of word meaning, including connotation (Garí Soler et al., 2022). We work with a corpus of two-sided debates which constitutes a scenario with interesting dynamics where we can find opinion disparity as well as concessions from either side. The application of an automatic coreference solver additionally allows us to work with different surface forms referring to the same entity. We carry out a qualitative and quantitative analysis of the proposed measures and investigate their usefulness in predicting a debate's outcome. Our measures reflect interesting word usage discrepancies between debate sides, and are directly applicable to other kinds of conversations.[2]

2 Related Work

2.1 Conceptual and Semantic Alignment

The first evidence of the tendency of speakers to align conceptually comes from Garrod and Anderson (1987) who noted that "once speakers have established a particular interpretation for an expression ... they try to avoid any potentially ambiguous use of that expression". Markman and Makin (1998) found that communication served to synchronize categorization (and thus to increase conceptual alignment): people who had worked together in a task involving toy construction pieces would sort pieces more similarly than two people who had collaborated on the task without talking.

Very few studies attempt to quantify conceptual alignment between dialog participants using automatic tools. Babcock et al. (2014); Ta et al. (2017) and Vrana et al. (2018) calculate the Latent Semantic Similarity (LSS, Landauer and Dumais 1997) between two speakers in a conversation. They find that LSS correlates positively with multiple dialog-level variables related to conversation length, expressive gestures or positive affect, among others. Xu (2021) uses more modern utterance representations derived from contextualized and static word representations (e.g., BERT (Devlin et al., 2019) and GloVe (Pennington et al., 2014)) to track utterance similarity throughout a dialog. The author finds patterns of global divergence and local convergence: semantic distance increases with temporal distance. These studies, however, compare the semantics of full utterances. We, instead, use contextualized word representations derived from BERT to compare how each side of a conversation uses a specific word. We partially follow work by Garí Soler et al. (2022), which compares word instance representations from sentences expressing opposing standpoints, and extend it to the two sides of a debate.

2.2 Asymmetric Alignment

We have so far described alignment as a mutual effort towards convergence, but one speaker can show more willingness to align than the other due to, among others, an asymmetry in their interpersonal relationship. For example, Danescu-Niculescu-Mizil et al. (2012) find that "higher-power" speakers (e.g., Wikipedia editors with Administrator status) receive more alignment (in terms of linguistic style markers used) than those of lower power. Xu et al. (2018), however, claim that this observation can be explained by low-level linguistic features such as utterance length, which tends to be larger in higher-power speakers and promotes a stronger alignment.

Asymmetric alignment has been observed in the context of debates, too. An electoral candidate's higher ranking in polls has been found to correlate with their convergence to the opponent's style (Romero et al., 2015) and the frequence with which the candidate manages to introduce or shift a topic (Prabhakaran et al., 2014). Similarly, Zhang et al. (2016) identify talking points of each side of a debate and investigate the extent to which each side talks about its own points or the opponent's points. They find that the winners tend to exhibit a drop in self-coverage, and are also more active in addressing the opponent's points.

In this study, we present both symmetric and asymmetric alignment measures. Relying on the same dataset as Zhang et al. (2016), we test the usefulness of asymmetric measures in predicting the winner of a debate.

[2] Our code is available at https://github.com/ainagari/LSalignment.

3 Data and Preprocessing

In this section we explain how we find the common vocabulary between debate sides and how we extract contextualized representations for words and phrases in this shared vocabulary.

3.1 Dataset

We use the Intelligence Squared Debates corpus (Zhang et al., 2016), IQ2, which contains 108 debates.[3] In each debate D there are two teams or sides ($S = \{f, a\}$), $f$or and $a$gainst the motion being discussed, made up of 2-3 people. Every debate has three parts: an introduction where each panelist is invited to present their main points in eight minutes; a 30-minute interactive part with questions from the moderator and the audience, and a conclusion where every participant has two minutes to make a closing statement. The audience casts a vote (for, against or undecided) before the debate and during the conclusion part. A team is considered to win a debate if it managed to "convert" more people, i.e., if the difference in the percentage of votes that their side received after vs before the debate is larger than that of the other team.

3.2 Shared Words

We are interested in observing the usage of words that are common to the two sides of a debate. We pos-tag and lemmatize[4] all the data. Following Garí Soler et al. (2022), we consider only nouns and verbs that are used at least three times by each side and for which all measures can be calculated.[5] We exclude stopwords and punctuation. We refer to the full shared vocabulary in a debate D as $V(D)$.

We additionally calculate tf-idf scores for every lemma, treating every debate as a document and determining the idf term from the whole dataset. We use these scores to select the most relevant and topic-specific words in a debate to be included in our analysis. See Table 5 (Appendix B) for examples of words ranked by tf-idf. Unless otherwise specified, we only use lemmas in $V(D)$ that are included in the top 200 by tf-idf ($V_{t200}(D)$). More information on the final vocabulary size used is given in Section 5.

3.3 Shared Entities

Coreference is a strongly present phenomenon in dialog, where speakers continously refer to already introduced entities with the use of pronouns, anaphoric expressions or paraphrases. Including chains of coreferent mentions in our analysis allows us to have a more complete and realistic picture of everything that is said about an entity, regardless of the way speakers refer to it. It also allows us to investigate the specific lexical choices made by each side, which may carry different connotations.[6]

We use the model presented by the UTD_NLP team (Li et al., 2022) at the recent CODI-CRAC 2022 shared task (Yu et al., 2022) which concerned anaphora phenomena in dialog. This was the best-performing coreference solver, with a 75.04 average CONLL F1 score on task 1 (identity anaphora resolution). We feed the model the full debates, including utterances by the moderator, the host and the audience. As a result we obtain coreference chains of terms referring to the same entity or concept.

We only include in our analysis those coreference chains with at least 3 co-referring terms uttered by each team. We observe that chains containing references to the panelists tend to contain errors, particularly when it comes to pronouns. This is understandable, as in a multi-party conversation it is not always clear who a speaker is referring to, especially from text alone. While it would be interesting to analyze how panelists talk about and refer to each other, we omit these chains from our analysis in order to reduce the errors due to automatic prediction.[7] After this filtering, we find an average of 16.3 coreference chains per debate, with an average length of 30.2[8] instances, which complete $V(D)$. We refer to this subset of the vocabulary as $V_C(D)$, and to the complementary subset made of lemmas as $V_W(D)$. Table 6 (Appendix B) shows examples of the coreference solver's output, which captures the use of synonyms, pronouns, phrases and paraphrases.

[3] Available with the convokit library (Chang et al., 2020).

[4] We use the nltk library.

[5] As explained in Section 4.2, certain measures have additional restrictions on the required number of instances.

[6] E.g., "Mexico's drug war" vs "America's drug war" as a way of emphasizing a party's responsibility or the war's reach or scope (example from the debate on "America Is To Blame For Mexico's Drug War").

[7] We automatically omit chains where one instance coincides with a panelist's full name, as well as all chains that are predominantly ($\geq 70\%$) made up of 1st and 2nd person pronouns.

[8] Counts do not include instances uttered by the host or the moderator.

3.4 Representation Extraction

Following Garí Soler et al. (2022), we extract contextualized representations for words and entities from BERT's (base, uncased) 10th layer. When a word is split into multiple tokens, we average the representations of each token. Since mentions in coreference chains can have multiple surface forms and BERT is sensitive to orthographic differences (Laicher et al., 2021), we additionally try using masking. We test different masking strategies to see which one yields representations that better reflect the differences in opinion between opposing sides. This experiment is detailed in Appendix A; as not masking gave the best result, all analyses presented in what follows are carried out without masking. We denote the set of instances of a word[9] $w \in V(D)$ uttered by a specific side $s \in S$ as $I_{w,s}$. We refer to the contextualized representation of an instance $i \in I_{w,s}$ as $\overrightarrow{i}$.

4 Alignment Measures

We propose measures which reflect different aspects of lexico-semantic alignment and compare them to lexical alignment measures used in previous work. We use a debate entitled "Don't Eat Anything With A Face"[10] as a running example to show the ranking of words obtained with each measure in Table 1. This table is to be discussed in more detail in Section 5.1. We compare the two sides of a debate, but our measures can be used to compare the word usages of two individual speakers.

We distinguish two main types of measures. With time-unaware (TU) measures, we compare word representations obtained from the debate as a whole, without taking into account the evolution or the change in word meaning as the debate progresses. Time-aware (TA) metrics, instead, explicitly compare representations at different temporal points of the debate. We make an additional distinction between symmetric and asymmetric measures. The former are measures of global or general alignment, whereas the latter are calculated separately for each side. We also consider measures of self-alignment, which quantify the semantic variation within a side.

Several of our measures rely on the averaged pairwise similarities ($psim$) between the representations of two sets of instances I and J (Equation 1). sim corresponds to a similarity measure. Unless otherwise specified, we use cosine similarity. It can be replaced with a distance measure, such as the Euclidean distance, in which case the results need to be interpreted accordingly.

$$psim(I, J) = \frac{\sum_{i \in I} \sum_{j \in J} sim(\overrightarrow{i}, \overrightarrow{j})}{|I| \times |J|} \tag{1}$$

4.1 Time-Unaware (TU) Measures

TU Self-Similarity (SS_{TU}) This metric measures the amount of variation that there is in the way one side of a debate uses one word. The SS_{TU} of a word w used by side s is calculated as the average pairwise similarity of instances within $I_{w,s}$:

$$SS_{TU}(w, s) = \frac{\sum_{i \in I_{w,s}} \sum_{i' \in I_{w,s}, i' \neq i} sim(\overrightarrow{i}, \overrightarrow{i'})}{|I_{w,s}|^2 - |I_{w,s}|} \tag{2}$$

With this metric we can examine the words that show the most and the least variation across sides (see Table 1). A global $SS_{TU}(s)$ measure for a side s of a debate D can be calculated by averaging the $SS_{TU}(s, w)$ of all words $w \in V(D)$.

TU Other-Similarity (OS_{TU}) This measure quantifies the similarity between the representation of a word w by each side in the debate. It gives an idea of how similar the meaning or the usage of a word is between the two sides.

$$OS_{TU}(w) = psim(I_{w,f}, I_{w,a}) \tag{3}$$

$OS_{TU}(w)$ allows us to see what words were the most and the least differently used between sides in the debate as a whole. We can calculate $OS_{TU}(D)$ for the whole debate by averaging the $OS_{TU}(w)$ of all words $w \in V(D)$.

Shared Vocabulary (SV) for a given concept We want to quantify the degree to which the two sides use the same surface forms to talk about the same thing. A given coreference chain $w \in V_C(D)$ consists of a set of instances uttered by either side (I_w). An instance $i \in I_w$ is realized with a specific surface form or realization r_i. The set of different realizations observed for chain w is denoted as R_w. To calculate this measure for a specific coreference chain ($SV(w)$), we first omit all mentions in I_w

[9] Here, a "word" is understood as a lemma with a specific PoS or as a concept described by a coreference chain. An "instance" is a specific usage of a word in context.

[10] This debate is clearly won by side FOR, which collects 21 additional votes after the debate, as opposed to AGAINST, which loses 8 votes.

that consist of a single pronoun. We only proceed if after this operation $|I_{w,s}| > 3$ for each side s.

We observe that chains often contain mentions that are very similar in form (e.g., *the war on drugs* vs *the drug war*). To avoid counting these as different realizations of the same concept, we perform a preliminary clustering of mentions in R_w based on their pairwise Levenshtein distance. Specifically, we merge realizations that are similar in form by means of hierarchical clustering with average linkage using a threshold of 5. After this step, expressions such as *this war* and *the war* are considered to be equivalent ways of referring to the concept expressed by w. Finally, we calculate the overlap between the two sides as follows:

$$SV(w) = \frac{\sum_{r \in R_w} min(|\{i \in I_{w,s} : r_i = r\}|, |\{j \in I_{w,s'} : r_j = r\}|)}{min(|I_{w,s}|, |I_{w,s'}|)} \quad (4)$$

$SV(w)$ ranges from 0 (no overlap) to 1 (maximum overlap). To be able to fairly compare the overlap of different coreference chains in the same debate, the score is normalized by the total number of instances involving concepts in $V_C(D)$:

$$SV(w, D) = SV(w) \frac{|I_w|}{\sum_{w' \in V_C(D)} |I_{w'}|} \quad (5)$$

This is the only measure that does not rely on contextualized representations. Despite the focus on surface form, we still consider it as a lexico-semantic measure because it is meant to be computed only on semantically equivalent expressions.

4.2 Time-Aware (TA) Measures

The metrics proposed here assume the existence of (at least) two time steps, an initial t_k and a posterior t_{k+1}. The set of instances of a word w by side s at time step k is denoted as $I_{w,s,k}$. We divide every debate into two halves (or time steps) following the number of tokens.[11] To calculate these measures for w, we require at least one instance of w per side and time step.

TA Self-Similarity (SS_{TA}) Analogously to SS_{TU}, this measure describes the self-variation of a word's usage within one side of the debate. SS_{TA}, however, takes time into account: we compare the representations at the beginning (t_k) and the end (t_{k+1}) of the debate to see if word usage has changed. While SS_{TU} represents the overall variation within one side, SS_{TA} captures evolution.

$$SS_{TA}(w, s) = psim(I_{w,s,k}, I_{w,s,k+1}) \quad (6)$$

Symmetric Approaching (*sApp*) This measure indicates whether the two sides came to use the word in a more similar way towards the end of the debate as opposed to the beginning. It is the difference in similarity between the two sides across the two time steps:

$$\begin{aligned} sApp(w) = psim(I_{w,f,k+1}, I_{w,a,k+1}) \\ - psim(I_{w,f,k}, I_{w,a,k}) \end{aligned} \quad (7)$$

A positive value means that representations of the two sides became closer by the end of the debate, compared to how they were at the beginning. Negative values indicate they grew further apart. The absolute value quantifies the magnitude of this difference.

Asymmetric Approaching (*asApp*) The measures introduced so far only tell us how close or similar representations are, or how much they approached each other. If the representations from the two sides are farther apart from each other at the end of the debate, what is the team that took the initiative of, or contributed the most to, this distancing? As explained in Section 2.2, Zhang et al. (2016) found that the winners of a debate tend to address the topics raised by their opponents. In a similar vein, we hypothesize that a side's initiative in approaching the other could be related to its outcome in the debate. To obtain a measure that reflects how much a side s has approached the other (s') in their usage of a word w, we take into consideration whether the representations by side s at t_{k+1} have come closer to the w representations from the other side s' at the previous time step:

$$\begin{aligned} sApp(w, s) = psim(I_{w,s,k+1}, I_{w,s',k}) \\ - psim(I_{w,s,k}, I_{w,s',k}) \end{aligned} \quad (8)$$

$asApp(w, s)$ is positive if the most recent word instances by side s ($I_{w,s,k+1}$) are closer in meaning to the initial instances of the word by the opposite side ($I_{w,s',k}$) than s's initial usage of w ($I_{w,s,k}$), and it is negative if they are farther away. We assume that the representations at time t_0 express the initial, unbiased meaning of a word by each side, whereas

[11] We considered the use of a sliding window, but discarded it due to data scarcity – most words do not occur in most windows. We leave the search for more sophisticated ways of taking temporality into account for future work.

			FOR	AGAINST
Time-Unaware	SS_{TU}	most similar	anything, farming, vegan, factory, attack	face, meat, farming, human, cancer
		least similar	life, grow, cow, die, study	attack, life, anything, die, study
	OS_{TU}	most similar	face, factory, meat, farming, cancer, human, vegetarian, vegan, animal, vegetable	
		least similar	life, attack, grow, die, study, cow, health, kill, fat, heart	
	SV	most overlap	fish, corn, plants, the globe / the world / the planet, vegetarians, face, cancer	
		least overlap	vitamin B12 / B12, the nation / the country, humans / human beings, this motion / the resolution	
Time-Aware	SS_{TA}	least evolved	anything, farming, vegan, factory, soil	face, cancer, cow, meat, human
		most evolved	grow, cow, life, die, kill	life, attack, study, health, die
	$sApp$	most approached	cow, grow, attack, anything, face, life, die, corn, meat, eat	
		most distanced	study, fat, vegetarian, health, soil, plant, food, farming, farm	
	$asApp$	most approached	cow, grow, face, human, attack	anything, attack, vegetable, eat, meat
		most distanced	vegetarian, fat, study, vegan, farming	study, health, food, plant, soil
	DS	common approaching	(+ balanced) factory, corn, attack, meat ... life, vegetable, cow, animal (- balanced)	
		common distancing	health, study, plant	
		opposite behavior	(+ extreme) diet$_f$, farm$_a$, food$_f$, farming$_a$ ... vegan$_a$, kill$_f$, grow$_f$, eat$_a$ (- extreme)	

Table 1: Word rankings obtained on the debate "Don't Eat Anything With A Face" by each measure: Time-unaware Self- and Other-Similarity (SS_{TU}, OS_{TU}), Shared Vocabulary (SV), Time-aware Self-Similarity (SS_{TA}), Symmetric and Asymmetric Approaching ($sApp$, $asApp$) and Driving Strength (DS). We use $V_{t200}(D)$ (28 words) (or $V_C(D)$ with 12 chains for SV). In DS (opposite behavior), subscripts indicate the side that approached.

representations at a posterior time step t_{k+1} reflect the evolution of the meaning of this word after having heard the other side. This measure indicates whether, and how much, the meaning of a word got closer to the pure, initial meaning of the word as presented by the other side. In this sense, it can capture the influence that the other side's statements may have had on s's representation of a word.

Driving Strength (DS) We combine the $asApp$ obtained by each side to obtain a normalized measure that indicates how much of the total approaching (or distancing) done by both sides each team is responsible for:

$$DS(w, s) = \frac{asApp(w, s)}{|asApp(w, s)|+|asApp(w, s')|} \tag{9}$$

$DS(w, s)$ can range between -1 and 1. Similarly to $asApp$, it is positive if s at t_{k+1} approached s' at t_k, and negative otherwise. For example, if $DS(w, s) = 0.5$ and $DS(w, s') = -0.5$, it means that both sides travelled the same distance, but s approached s' and s' got farther away from s. In this case, $sApp(w)$ would be 0, which would not reflect the fact that one side approached the other.

To sum up, we have three symmetric measures, two time-unaware (OS_{TU}, SV) and one time-aware ($sApp$); and four asymmetric measures, one time-unaware (SS_{TU}) and three time-aware (SS_{TA}, $asApp$ and DS). See Figure 2 in Appendix B for an illustration of how each measure behaves in different situations.

4.3 Lexical Measures

We calculate a series of measures available from the Dialign software (Dubuisson Duplessis et al., 2021) which take into account different aspects of lexical alignment (amount of self-/other-repetition, variety of expressions, complexity of lexical patterns, orientation of alignment...). We provide a list of the metrics in Appendix C. A more thorough description can be found in Dubuisson Duplessis et al. (2021). We include these measures to investigate the correlation between lexical and lexico-semantic alignment, and to combine them with our proposed measures for predicting a debate's winning side.

5 Analysis

In Section 5.1, we carry out a qualitative analysis of the kinds of phenomena our measures reflect. We do so following our running example and looking at the results for individual words presented in Table 1. Section 5.2 investigates the measures' behavior when calculated at the dataset level.

The vocabulary used for the SV metric is $V_C(D)$. For all other metrics, we use word lemmas from $V_{t200}(D)$ provided that at least one instance is available for each time step and side.[12] This consists of 33 lemmas on average.

[12] This restriction is not necessary for time-unaware metrics, but we apply it so the same vocabulary is used across all measures.

5.1 Word-level Analysis

We find that our measures, calculated with BERT, capture a wide range of usage phenomena. Apart from differences in word sense (WS) and connotation (CN), they are also sensitive to unusual word usages or expressions (U), to differences in collocations or subject/object preferences (CL), and to the distinction between entities and common nouns (E). We present several examples below.

In Table 1, we can see that the noun *attack* has one of the lowest OS_{TU}. This reflects the fact that FOR talks exclusively about heart attacks related to meat consumption, whereas AGAINST also mentions panic attacks (due to a worse mental health presumably caused by veganism) and attacks in a metaphorical sense ("Being vegan is an attack on the poor") **(WS)**. This also explains why SS_{TU}(*attack*, *f*) is quite high. Another word with low OS_{TU} is *die*: while AGAINST talks more often about animals dying, FOR also mentions people dying from diseases related to elevated meat consumption **(CL)**. *Factory*, instead, has a high OS_{TU}, and it is used by both sides almost exclusively in the context of "factory farm" **(CL)**.

Farming displays a very high SS_{TU} for FOR. This is because its instances almost exclusively contain criticism to factory farming (e.g., "factory farming is an abomination", "factory farming is bad") **(CN, CL)**. *Life*, instead, is one of the words with highest variation within both sides of the debate. Both FOR and AGAINST indeed make a varied use of this word: to talk about animals' or humans' life, to talk about killing ("taking someone's life"), about health ("life expectancy"), or to refer to "aliveness" in general ("life often comes from death") **(WS, CL)**.

When it comes to *sApp*, we find that *vegetarian* is among the words that became most distant between sides. This is because in the debate, AGAINST starts talking about their failed past as a vegetarian and the benefits that they expected from it. But beyond that, instances of *vegetarian* by each side occur in sentences that highlight the benefits of the dietary choice (meat-based vs vegetarian) that is being defended or the problems created by the opposing side's choice **(CN)**. In the case of *study*, also with low $sApp$, FOR focuses on a specific study called "the China study" during the second half, whereas in the rest of the debate both sides bring up multiple studies in a similar way **(E, CL)**.

Cow and *grow*, instead, are two of the words whose representation becomes most similar. FOR uses *cow* in the expression "Holy cow" in the first part of the debate, but its subsequent usages are literal (i.e., not idiomatic), like those by AGAINST **(U)**. *Grow* is used with the meaning of "growing up" by FOR in the first half, while in the rest of the debate it tends to be used in the sense of growing crops **(WS)**. *Anything* and *face*, both with high $sApp$, are two words included in the title of the motion, which is repeated multiple times throughout the debate **(CL)**. However, *face* is initially used by FOR to talk about empathy when looking into someone else's face, which explains the high value of $asApp(face, f)$. The case of *corn* is also interesting: its high $sApp$ can be attributed to an unusual usage of *corn* by AGAINST in the first half ("corn has ears"), to refer to the fact that plants are sentient. FOR picks up on this on the second half of the debate ("not one ear of that corn is going to be eaten") **(U)**.

Looking at the coreference chains and their shared vocabulary SV between sides, we do not observe anything particularly controversial in this debate. When talking about humans, AGAINST uses mostly *humans*, pronouns (*we* and *our*, which are not taken into account in our measure) or, in one occasion, *human beings*. FOR uses also *mankind*, *man* and *people*. FOR very often refers to vitamin B12 simply as *B12*, whereas AGAINST uses the whole phrase.

We also calculate the correlation between our measures and word frequency, counted as the number of occurrences of a word in a debate as a whole (for symmetric measures) and by side (for asymmetric ones). Results show that none of our measures is affected by frequency ($|\rho| < 0.04$).

5.2 Dataset-level Analysis

In Table 2, we present the descriptive statistics of the measures as calculated on the whole collection of debates. Values obtained relying on Euclidean distance are included in Appendix B. Similarly to Garí Soler et al. (2022), we observe that measures that directly reflect similarity (OS_{TU}, SS_{TU} and SS_{TA}) have high values in a narrow range, due the anisotropy of BERT representations (Ethayarajh, 2019). For the same reason, measures that subtract two similarities ($sApp$ and $asApp$) have very low values. As expected in a debate setting, we find that other-similarity (OS_{TU}) is overall slightly lower than self-similarity measures (SS_{TU} and SS_{TA},

Measure	Avg	Min	Max	Std
SS_{TU}	0.71	0.63	0.75	0.02
OS_{TU}	0.69	0.62	0.72	0.02
SS_{TA}	0.70	0.61	0.75	0.02
$sApp$	0.01	-0.05	0.06	0.02
$asApp$	0.00	-0.04	0.05	0.01
DS	0.02	-0.25	0.34	0.11
SV	0.88	0.64	1.0	0.07

Table 2: Descriptive statistics of the proposed measures calculated on IQ2 with V_{t200} (or V_C for SV).

$p < 0.05$),[13] which indicates that a side's usage of a word tends to be more stable and coherent than usages across sides. The mean values of $sApp$ and $asApp$ are almost 0, suggesting that, on the whole, sides do not really tend to come closer to each other by the end of the debate in terms of word usage.

We also calculate the inter-correlations between our measures.[14] The only strong correlations found ($\rho > 0.5, p < 0.001$) are between SS_{TU} and SS_{TA} (0.93); and between $asApp$ and DS (0.77). This is not surprising, as these measures are related by definition. While each measure is contributing a specific kind of information, SS_{TA} could probably benefit from a different treatment of temporality. Correlations with Dialign measures are all weak ($\rho < 0.31$). This suggests that lexical and lexico-semantic alignment do not necessarily come together. This makes sense in a debate setting, where we expect semantic divergence on a very specific topic; but this result could be different in other types of conversations.

We compare the values of our asymmetric measures (SS_{TU}, SS_{TA}, $asApp$ and DS) when different sides win the debate. We use the 105 debates that do not end in a tie (52 where FOR wins, 53 where AGAINST wins).[15] We only find significant ($p < 0.05$) differences with the SS_{TA} measure. However, both $SS_{TA}(f)$ and $SS_{TA}(a)$ are overall slightly higher when AGAINST wins. Therefore, we cannot conclude that, when taken individually, the proposed measures reflect the winning side of a debate.

6 Toward Automatic Winner Prediction

We investigate whether the proposed measures can be used in combination in a supervised classification setting to automatically predict the winning side of a debate. For this experiment we again use the 105 debates where one side won. Given the little data available, we obtain model predictions in a leave-one-out setting. We fit a logistic regression model using different sets of features.

Measures	sim/dist	vocab.	Accuracy
asOurs	cos	V	**0.57**
asOurs	cos	V_{t200+C}	**0.57**
asOurs	eucl	V_{t200+C}	**0.57**
asOurs	eucl	V_{t200}	**0.57**
Ours	eucl	V_{t200}	**0.57**
asAll	cos	V_{t200+C}	0.54
asDia	-	-	0.52
Majority class baseline			0.50
Length baseline			0.49

Table 3: Results of different models on the winner prediction task. We include the best result obtained with each individual parameter.

Features We use three sets of asymmetric measures (calculated for each side): ours ($asOurs$), dialign measures ($asDia$), and all of them combined ($asAll$). Additionally, we try using our symmetric and asymmetric measures in combination ($Ours$). We experiment with different parameters when calculating our measures. We use cosine similarity (*cos*) or euclidean distance (*eucl*) and different vocabularies: everything (V) or words that are within the 200 words with highest tf-idf (V_{t200}), optionally in combination with V_C (V_{t200+C}).

Results Table 3 presents a summary of the results, in terms of accuracy, including the models that obtained the highest scores and at least one result (the best) for each parameter value. We also show the results of a majority class baseline that always predicts the class AGAINST (the most common in IQ2) as well as of a model that only relies on simple length-related features (*Length*).[16] The complete results can be found in Appendix B. Our asymmetric measures on their own obtain the best result (0.57) relying on different combinations of similarity or distance metrics and vocabularies. The same result is achieved with all our metrics calculated with Euclidean distance and V_{t200}. We do not observe a clear pattern as to the best similarity/distance or vocabulary to use. The combination of our measures with dialign or with our

[13] Determined with Mann Whitney U tests.

[14] We do not mix symmetric with asymmetric measures.

[15] We run t-tests or Mann Whitney U's tests according to normality, which is determined with Shapiro-Wilk tests.

[16] The following dialign measures: Num. utterances, num. tokens, % of tokens per side.

symmetric measures does not provide an advantage (0.54). Comparing to the best results obtained by the Dialign measures on their own (0.52), we conclude that asymmetric lexico-semantic measures are more useful for predicting a debate's winning side.

Most results are superior to the baselines, although not by a very large margin. This highlights both the importance of parameter optimization as well as the difficulty of the task. Predicting the winning side of a debate is hard, even for humans. Accuracy is below that obtained by Zhang et al. (2016) using conversational flow features in a similar setting (0.65). Overall, these results show that our asymmetric measures can, when used in combination, assist in (but not solve) this task.

7 Conclusion and Future Work

We have introduced and discussed the notion of lexico-semantic alignment. We have proposed a set of measures relying on contextualized word representations which are designed to account for different aspects of alignment, such as temporality and asymmetry. Our qualitative analysis shows that our metrics calculated with BERT reflect multiple semantic phenomena (e.g., collocations, connotation) that characterize the way each side of a debate uses specific words. We have also shown that the debate-level information provided by these metrics can be helpful for predicting a debate's winner.

In future work, we plan to study our measures' behavior on other kinds of conversations where the focus would be on individual speakers, such as task-oriented dialogs or everyday conversations involving multiple topics. We think that they are also potentially useful for detecting cases of misunderstanding due to lexical ambiguity or due to a language proficiency level mismatch between interlocutors. We can also refine our measures with a more fine-grained treatment of temporality and including information of the speaker who introduced each word. Finally, an obvious extension would be to experiment with different representations, e.g., from other language models.

Limitations

Coreference resolution quality. While we have taken care of choosing a good coreference solver and filtering out chains referring to speakers, the automatic resolution of coreference in dialog remains a challenging task. The quality of the tool has a direct impact on our SV measure, but also on our other estimations when including coreference chains.

The lack of manual annotation for lexico-semantic alignment makes it hard to run a systematic evaluation of the quality of the proposed measures. Our qualitative analysis provides valuable insight, but on one debate only. The classifier experiments demonstrate their usefulness for winner prediction, but they do not constitute an intrinsic evaluation. However, we note that annotating conversations with such information is bound to be a highly subjective, challenging and expensive task.

Acknowledgements

We thank Francesc Garí Soler for the interesting discussions, Shengjie Li for sharing the coreference resolution model, and the anonymous reviewers for their helpful comments and suggestions. This research is supported by the Télécom Paris research chair on Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS) and by the Agence Nationale de la Recherche, REVITALISE project (ANR-21-CE33-0016).

References

Meghan J Babcock, Vivian P Ta, and William Ickes. 2014. Latent semantic similarity and language style matching in initial dyadic interactions. *Journal of Language and Social Psychology*, 33(1):78–88.

Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Alexandra A Cleland and Martin J Pickering. 2003. The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, 49(2):214–230.

Frederick G Conrad and Michael F Schober. 2000. Clarifying question meaning in a household telephone survey. *Public opinion quarterly*, 64(1):1–28.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Guillaume Dubuisson Duplessis, Caroline Langlet, Chloé Clavel, and Frédéric Landragin. 2021. Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. *Language Resources and Evaluation*, 55:353–388.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2022. One word, two sides: Traces of stance in contextualized word representations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3950–3959, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Thomas K Landauer and Susan T Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological review*, 104(2):211.

Shengjie Li, Hideo Kobayashi, and Vincent Ng. 2022. Neural anaphora resolution in dialogue revisited. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 32–47, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Arthur B Markman and Valerie S Makin. 1998. Referential communication and category acquisition. *Journal of Experimental Psychology: General*, 127(4):331.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL-08: HLT, Short Papers*, pages 169–172, Columbus, Ohio. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An indicator of power in political debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1481–1486, Doha, Qatar. Association for Computational Linguistics.

David Reitter and Johanna D. Moore. 2007. Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815, Prague, Czech Republic. Association for Computational Linguistics.

Daniel Richardson, Rick Dale, and Kevin Shockley. 2008. Synchrony and swing in conversation: Coordination, temporal dynamics, and communication. *Embodied communication in humans and machines*, pages 75–94.

Daniel M Romero, Roderick I Swaab, Brian Uzzi, and Adam D Galinsky. 2015. Mimicry is presidential: Linguistic style matching in presidential debates and improved polling numbers. *Personality and Social Psychology Bulletin*, 41(10):1311–1319.

Michael F Schober. 2005. Conceptual Alignment in Conversation. *Other minds: How humans bridge the divide between self and others*, pages 239–252.

Kevin Shockley, Marie-Vee Santana, and Carol A Fowler. 2003. Mutual Interpersonal postural Constraints are Involved in Cooperative Conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2):326.

Arjen Stolk, Lennart Verhagen, and Ivan Toni. 2016. Conceptual Alignment: How Brains Achieve Mutual Understanding. *Trends in cognitive sciences*, 20(3):180–191.

Richard L Street Jr. 1984. Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, 11(2):139–169.

Vivian P Ta, Meghan J Babcock, and William Ickes. 2017. Developing Latent Semantic Similarity in Initial, Unstructured Interactions: The Words May Be All You Need. *Journal of Language and Social Psychology*, 36(2):143–166.

Scott R Vrana, Dylan T Vrana, Louis A Penner, Susan Eggly, Richard B Slatcher, and Nao Hagiwara. 2018. Latent Semantic Analysis: A new measure of patient-physician communication. *Social Science & Medicine*, 198:22–26.

Yang Xu. 2021. Global divergence and local convergence of utterance semantic representations in dialogue. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 116–124, Online. Association for Computational Linguistics.

Yang Xu, Jeremy Cole, and David Reitter. 2018. Not that much power: Linguistic alignment is influenced more by low-level linguistic features rather than social power. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 601–610, Melbourne, Australia. Association for Computational Linguistics.

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2022. The CODI-CRAC 2022 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–14, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

A Masking Experiment

We experiment with three different masking strategies: we replace the target word or phrase with a single [MASK] token (*one-mask*) or with as many [MASK] tokens as the original number of subwords (*multi-mask*) and compare these to the default approach of not masking (*no-mask*). The goal of masking is to abstract away from the surface form chosen by the speaker to refer to a concept, keeping only the meaning provided by the context in which it is used.

	no-mask	**one-mask**	**multi-mask**
V_W	**0.70**	0.69	0.69
V_C	**0.75**	0.71	0.71

Table 4: Accuracy of the three masking strategies with different kinds of shared vocabulary terms.

In order to find the best masking strategy to extract BERT representations for common words and concepts, we run a similar experiment to Garí Soler et al. (2022). Our evaluation criterion is the following: since we know that the two sides of a debate have opposing opinions, we want word representations found in one side to be more similar to each other than to representations from the other side. In other words, we expect the WITHIN-side similarity to be higher than the BETWEEN-side similarity. We verify which of the masking strategies yields representations that most clearly reflect the difference in opinion.

To obtain the data for a word w in a debate D, we randomly split the instances of a given side $I_{w,s}$ into two equally-sized sets of size ≥ 3, when possible. This results in four sentence sets (FOR$_1$, FOR$_2$, AGAINST$_1$, AGAINST$_2$). We obtain a word representation from each sentence set by averaging the contextualized representations of all word instances in it. With this data, we can run four comparisons: WITHIN-FOR, WITHIN-AGAINST, BETWEEN-1 (with FOR$_1$ and AGAINST$_1$,) and BETWEEN-2. We calculate the cosine similarity for each of these comparisons.

Accuracy is calculated as the proportion of (WITHIN, BETWEEN) comparison pairs (four per word) where the BETWEEN comparison had a lower similarity. Our experiments on V_W involve a total of 4,965 words (an average of 46 words per debate), which amount to 19,860 comparison pairs. For those on V_C, 841 concepts are used (an average of 7.8 per debate and a total of 3,364 comparison pairs).

Results are presented in Table 4 separately for common lemmas (V_W) and for concepts in coreference chains (V_C). Accuracy is higher in the *no-mask* setting, for both kinds of vocabulary elements, but particularly so for concepts found in coreference chains. We also note that accuracy is lower than in Garí Soler et al.'s 2022 experiments. This is not surprising, however, as they used sentences explicitly expressing a stance, while in debates not all sentences express an opinion unequivocally.

B Additional Tables and Figures

- Table 5: examples of words ranked by tf-idf.

Abolish the dead penalty

Top: penalty, death, abolish, parole, prison, punishment, deterrence, execution, sentence, victim...

Bottom: ...provide, learn, opening, university, week, city, work, open, power, turn

---

Global warming is not a crisis

Top: warming, climate, warm, temperature, greenhouse, crisis, atmosphere, dioxide, scientist, CO2...

Bottom: ...school, spend, friend, pay, set, week, city, everyone, view, lose

Table 5: Top and bottom noun and verb lemmas extracted from two debates ranked by tf-idf. Proper nouns are omitted.

- Table 6: examples of the coreference solver's output.
- Figure 2: illustration of the measures' behaviour on different toy examples.
- Table 7: descriptive statistics of our measures calculated with Euclidean distance.
- Table 8: Results of all tested settings on debate's winner prediction.

C Dialign Measures

We present below the list of Dialign measures (Dubuisson Duplessis et al., 2021) used in the paper. Note that the software finds matching lexical patterns in the conversation which can consist of multiple tokens; these are referred to as "expressions".

Symmetric (speaker-independent) measures:

- **Number of utterances**
- **Number of tokens**
- **Expression Lexicon Size (ELS)**† : number of established expressions in the dialog.
- **Expression Variety (EV)**† : variety of the shared expression lexicon.
- **Expression Repetition (ER)**† : proportion of tokens dedicated to repetitions.
- **Vocabulary overlap**† : ratio of shared vocabulary items.
- **ENTR**† : entropy of the lengths (in tokens) of shared expressions.
- **L**† : average length of shared expressions.
- **LMAX**† : maximum length of shared expressions.

The symmetric measures marked with † also have an asymmetric (speaker-dependent) version. Other asymmetric measures are:

- **Tokens (%)**
- **Initiated Expression**: ratio of shared expressions initiated by a speaker.

Debate title	Coreference examples
Obesity Is The Government's Business	We were also concerned about what was happening in **children**. For every **kid**, they get a report card that doesn't just give their arithmetic score.
	We cover some 8 percent of the **U.S.** work force for long term disability (...) (...) the surgeon of the general of the **United States** raised the alarm about (...) And **America** wouldn't be going broke.
Too Many Kids Go To College	(...) going to college is part of **the American dream** (...) We need to do better, and we can't give up on **the American dream**.
	(...) Students in the first tier system and a whole lot of **very expensive elite colleges** (...) (...) that is true of **the elite universities**.
The President Has Exceeded His Constitutional Authority by Waging War Without Congressional Authorization	(...) air strikes on **ISIS** (...) (...) the **Islamic State** didn't exist in 2001 (...) (...) **it** has distanced **itself** from the core al-Qaeda leadership (...)

Table 6: Examples of the coreference solver's output for different debates. We find coreference chains containing synonyms, phrases, paraphrases and pronouns.

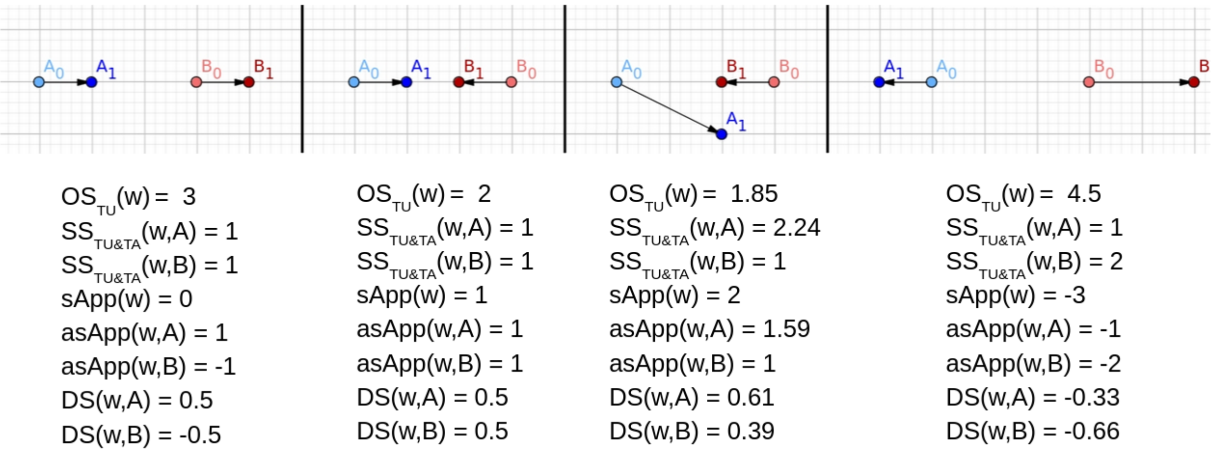

Figure 2: Values obtained with each measure in different toy situations. For ease of interpretation, we calculate the measures with Euclidean distance. Values of SS_{TU}, SS_{TA} and OS_{TU} are to be interpreted as distances. The sign in measures that rely on similarity differences ($sApp$, $asApp$, DS) has been adapted so a negative value indicates distancing. A and B represent the two sides of a debate, and subscripts 0 and 1 refer to the two time steps. In such simplified setting, with only two instances per side, SS_{TU} and SS_{TA} are equivalent.

Measure	**Avg**	**Min**	**Max**	**Std**
SS_{TU}	14.92	13.29	17.37	0.64
OS_{TU}	15.52	14.33	17.52	0.56
SS_{TA}	15.29	13.74	17.83	0.66
$sApp$	0.167	-1.51	1.58	0.45
$asApp$	0.00	-1.16	1.03	0.32
DS	0.01	-0.23	0.32	0.12

Table 7: Descriptive statistics of the proposed measures calculated on IQ2 using Euclidean distance and V_{t200}.

Measures	**sim/dist**	**vocab.**	**Accuracy**
asOurs	cos	V	**0.57**
asOurs	cos	V_{t200+C}	**0.57**
asOurs	eucl	V_{t200+C}	**0.57**
asOurs	eucl	V_{t200}	**0.57**
Ours	eucl	V_{t200}	**0.57**
Ours	cos	V_{t200}	0.55
asOurs	cos	V_{t200}	0.54
asAll	cos	V_{t200+C}	0.54
Ours	eucl	V_{t200+C}	0.54
asAll	eucl	V_{t200+C}	0.54
asDia	-	-	0.52
asAll	cos	V	0.52
asOurs	eucl	V	0.52
asAll	eucl	V	0.52
asAll	eucl	V_{t200}	0.51
Ours	cos	V_{t200+C}	0.50
asAll	cos	V_{t200}	0.50
Ours	cos	V	0.49
Ours	eucl	V	0.47
Majority class baseline			0.50
Length baseline			0.49

Table 8: Complete results on the debate's winner prediction task.