Towards a model of conversational coherence: its potential for the automatic evaluation of dialogue systems

Alexandra Burchill
University of Cambridge

April 13, 2019

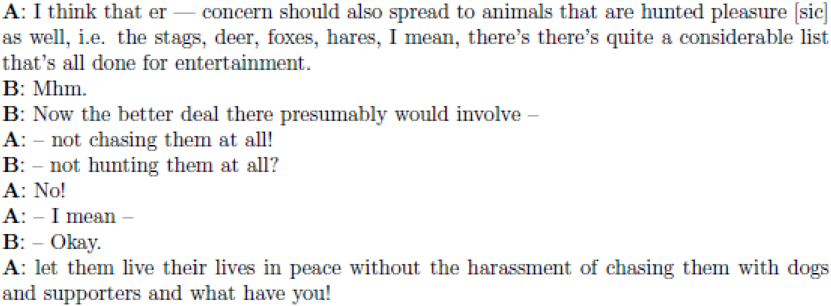
Overview

- Dialogue Systems
- Coherence in discourse and conversation
- Methods for evaluating dialogue system output
- The potential of coherence as an automatic evaluation metric for dialogue systems

Dialogue Systems

- Generate output in response to user input, i.e. systems which talk to users
- Conversations between a dialogue system and its user(s) may or may not be goal-oriented, e.g. flight booking agent vs. ‘chatbot’
- Have many current and potential applications, within: industry (Google, Alexa, Siri), research, and language learning (as a tool for second language learners to practice conversational skills)
- No current dialogue system is wholly human-like in its conversational abilities, concerning the scope of topics and qualitative properties of output

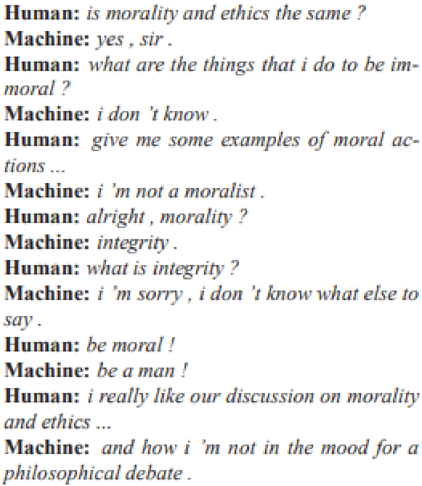
Human-human conversation



A: I think that er — concern should also spread to animals that are hunted pleasure [sic] as well, i.e. the stags, deer, foxes, hares, I mean, there's there's quite a considerable list that's all done for entertainment.
B: Mhm.
B: Now the better deal there presumably would involve –
A: – not chasing them at all!
B: – not hunting them at all?
A: No!
A: – I mean –
B: – Okay.
A: let them live their lives in peace without the harassment of chasing them with dogs and supporters and what have you!

Figure 1: extract from human-human conversation, on the topic of animal rights. British National Corpus (2007, FLH: 35–43)

Human-machine conversation



Human: *is morality and ethics the same ?*
Machine: *yes , sir .*
Human: *what are the things that i do to be immoral ?*
Machine: *i don 't know .*
Human: *give me some examples of moral actions ...*
Machine: *i 'm not a moralist .*
Human: *alright , morality ?*
Machine: *integrity .*
Human: *what is integrity ?*
Machine: *i 'm sorry , i don 't know what else to say .*
Human: *be moral !*
Machine: *be a man !*
Human: *i really like our discussion on morality and ethics ...*
Machine: *and how i 'm not in the mood for a philosophical debate .*

Figure 2: human-machine conversation with a neural conversational model on the topic of morality. From: Vinyals, O. and Q. Le (2015). A neural conversational model". In: arXiv preprint arXiv:1506.05869.

Unsupervised dialogue systems

- A variety of dialogue systems are based on Neural Networks (encoder-decoder) trained using unsupervised learning algorithms - many different models!
    - Neural Network: artificial network of ‘neurons’ - small processing units - organised into ‘input’ and ‘output’ layers, with ‘hidden’ layers inbetween
    - ‘encoder-decoder’: input utterances are encoded as vectors (representing the properties of input), which are passed through hidden layers of units, to obtain an output vector, which is decoded into the most probable output response utterance
- ‘Unsupervised’: the model consists of a Neural Network, which learns the statistical properties of data it is trained on, without any labels on training data to help it learn the properties of the data.

Neural Networks

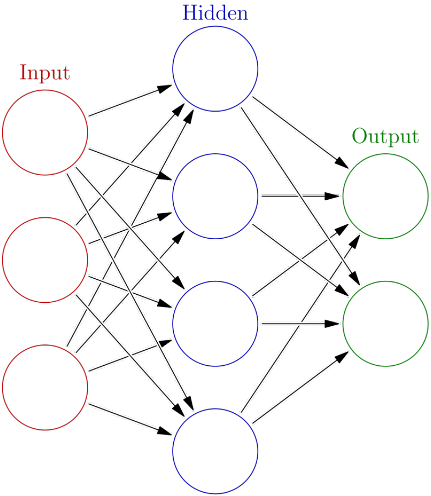


Figure 3: https://commons.wikimedia.org/w/index.php?curid=24913461

Motivations for an investigation of conversational coherence

Dialogue systems:

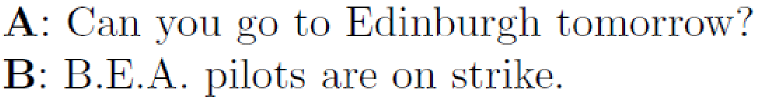
- Have many applications within industry, research, and as potential tools for learners of a language to gain conversational experience
- Currently lack automatic evaluation methods for their improvement, or any standardised procedures for evaluating performance
- Could benefit from evaluation, as this may identify areas of improvement: qualities of training data, training algorithm, architecture of dialogue system, etc.

We should aim to make evaluation as objective a procedure as possible. Automatically obtaining scores, rather than relying on human judgements, would be a step towards this.

What is coherence?

- A property of successful communication and discourse
- But can it be quantified for the purposes of automatic evaluation?

Where local coherence holds between two adjacent utterances, an immediate communicative goal is satisfied, i.e. providing the relevant information requested by a conversational partner. But are there any linguistic properties which determine what a locally coherent response will consist of?



A: Can you go to Edinburgh tomorrow?
B: B.E.A. pilots are on strike.

Figure 4: Widdowson (1979: 96)

Existing theories of coherence and discourse structure

Existing computational theories of discourse coherence are based on monologic texts:

- Rhetorical Structure Theory (Mann & Thompson 1987): a relation-based framework for coherence.
- Centering Framework (Grosz. et al 1995): an entity–based framework for coherence.

Do these provide a suitable theoretical basis for developing models for automatically evaluating the extent to which a conversation (i.e. dialogue) exhibits coherence?

Rhetorical Structure Theory

A framework for modelling the organisation of text. A coherent discourse or text can be represented as a tree-like structure, where spans of text are in asymmetric relations with each other (nucleus–satellite). A coherent discourse will have all spans of text contained within a hierarchical tree–structure, whose constituents are spans of text.

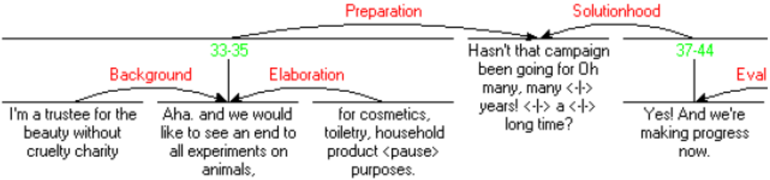

Preparation
Solutionhood
33-35
Background
Elaboration
Hasn't that campaign been going for Oh many, many <-|-> years! <-|-> a <-|-> long time?
37-44
Eval
I'm a trustee for the beauty without cruelty charity
Aha. and we would like to see an end to all experiments on animals,
for cosmetics, toiletry, household product <pause> purposes.
Yes! And we're making progress now.

Figure 5: part of a hierarchical representation of a discourse, whereby spans of text are linked by the rhetorical and structural relations which are present.

Centering Framework

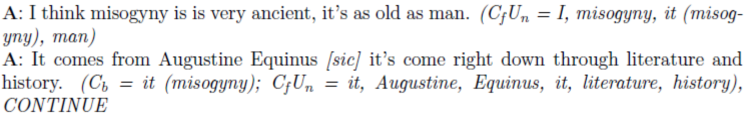
An entity-based approach to modelling coherence in discourse. The focus/topic of a discourse may shift as it progresses, and capturing the transitions between focal entities in discourse, across utterances in sequence, can model how coherence is exhibited locally in discourse.

- ‘Centers’ are the entities within each utterance
- Each utterance has a set of ‘forwards–looking’ centers (the entities in the sentence) and a single ‘backwards–looking’ center, which is selected from the set of forwards–looking centers of the preceding utterance, to serve as the focus of attention in the current utterance.
- Apart from the initial utterance, each utterance in a discourse will have a set of forwards–looking centers, and a backwards–looking center.

Centering Framework

4 types of transitions: continue, retain, smooth shift, rough shift.

- Continue/retain: an utterance, and the utterance which immediately follows within the discourse, have the same backwards–looking center (focus of attention).
- Shift: A change in the center between two adjacent utterances.
- Forward–looking centers are ranked, according to syntactic roles, and semantic, pragmatic and discourse/functional criteria.



A: I think misogyny is is very ancient, it's as old as man. *($C_fU_n = I$, misogyny, it (misogyny), man)*

A: It comes from Augustine Equinus *[sic]* it's come right down through literature and history. *(C_b = it (misogyny); C_fU_n = it, Augustine, Equinus, it, literature, history), CONTINUE*

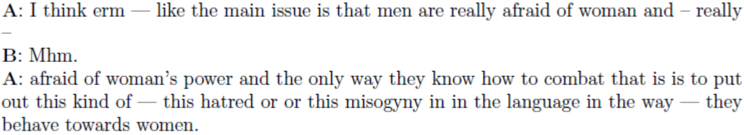
Figure 6: centers and 'continue' transition between two successive utterances in a conversation (BNC 2007, FL7: 18–19)

Problems with applying the Centering Framework and Rhetorical Structure relations to conversations

- Language is used differently by speakers within conversations, as opposed to its use in written texts
- Conversations are constructed differently - goals of discourse are negotiated by speakers, and turns (units of analysis) may overlap.
- Coherence, and the organisation and goals of discourse as they occur in conversation, are dynamically determined.
- Conversation consists of shared actions
- Need to annotate data – Part–of–Speech tagging, syntactic (and semantic) parsing of conversations, before entity or relation–based evaluation of coherence could take place.

Identifying centers

- Discourse markers & phrases: speakers use utterances to situate their contributions to a conversation.
- This helps enable other participants to interpret and understand their utterances, but as features of the utterance, should not be analysed as centers by the centering algorithm.
- How could discourse markers, and phrases with functions pertaining to the organisation of discourse, be recognised by an automatic model of coherence?



A: I think erm — like the main issue is that men are really afraid of woman and – really –

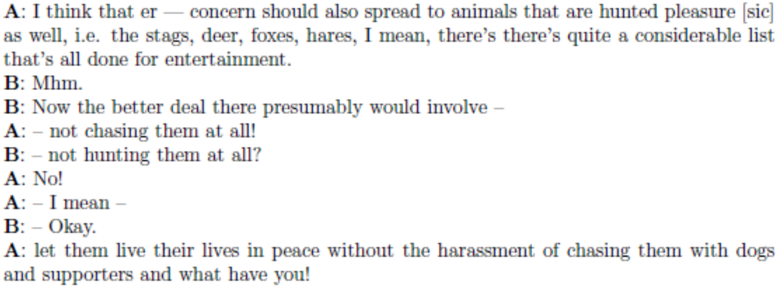
B: Mhm.

A: afraid of woman's power and the only way they know how to combat that is is to put out this kind of — this hatred or or this misogyny in in the language in the way — they behave towards women.

Figure 7: 'I' in 'I think' and 'the main issue' are parts of phrases which frame the utterance as a contribution to the discourse. These would be tagged as NPs, but should not be identified as centers of the discourse (BNC 2007, FL7: 25–27)

Overlapping turns & rhetorical relations

The human–human conversation we saw earlier has some overlap between turns, whereby speakers negotiate topic as they attempt to predict the conversational intentions of each other. The units of discourse, in terms of their function, are, at times, discontinuous in nature (i.e. spanning across multiple utterances/turns).



A: I think that er — concern should also spread to animals that are hunted pleasure [sic] as well, i.e. the stags, deer, foxes, hares, I mean, there's there's quite a considerable list that's all done for entertainment.
B: Mhm.
B: Now the better deal there presumably would involve –
A: – not chasing them at all!
B: – not hunting them at all?
A: No!
A: – I mean –
B: – Okay.
A: let them live their lives in peace without the harassment of chasing them with dogs and supporters and what have you!

Can coherence be quantified?

Can we score conversations on the degree of coherence that they exhibit?

- The distribution (spread & range) of centering transitions across the discourse could provide a calculable basis for a score of coherence.
- Some labelling of utterance functions/rhetorical relations may be necessary to determine how appropriate/coherent entity transitions are at a given point in the discourse.
- Annotating conversations for rhetorical relations - how could this be implemented automatically?
- The nature of coherence in any given conversation/discourse may be exhibited differently according to the genre of the conversation. Pragmatic and sociolinguistic factors would therefore need to be taken into account when modelling coherence, with different methods of evaluation used accordingly.

Conclusions

- Conversational analysis can help us identify how coherence is achieved in conversation, which we should aim to achieve in the design of dialogue systems.
- Existing computational theories of discourse coherence need to be adapted, to provide a basis for developing automatic procedures for coherence–based evaluation.
- Modelling transitions between discourse entities, and the rhetorical structures of conversations, may provide a starting point for evaluating dialogue system performance.