T-(538|725)-MALV, Natural Language Processing Information Structure and Newness of Information

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]School of Computer Science, Reykjavik University, Iceland

November 2009

Outline

Outline

Discourse Analysis (review)

The purpose of Discourse Analysis

To find patterns in regular language use that describe how **Discourse Devices** (í. orðræðuáhöld) are used to achieve various **Discourse Functions** (í. orðræðumarkmiðum).

Machines can use this to...

- ...process and interpret naturally occurring text.
- ...interact with humans on their terms.
- ...assist humans with communicating with each other.

Discourse Analysis (review)

Context and Discourse Model

- **Discourse Model** (í. orðræðulíkan) keeps track of the **Context** (í. samhengi) of the ongoing discourse.
- The **Textual Context** is represented by a growing set of **Discourse Entities** that correspond to what has been discussed so far in the text.
- The correct interpretation of the discourse depends heavily on this context, since **Referring Expressions** may refer to it at any time.

Outline

The Linearisation Problem

Sentence by Sentence Construction and Local Context

- Discourse is produced sentence by sentence in a linear fashion.
- The text of a sentence provides the immediate context (co-text) for the following sentence.
- This local context affects the interpretation.

Example

- "I saw a ghost. It was real!"

The Linearisation Problem

The Local Context and Interpretation

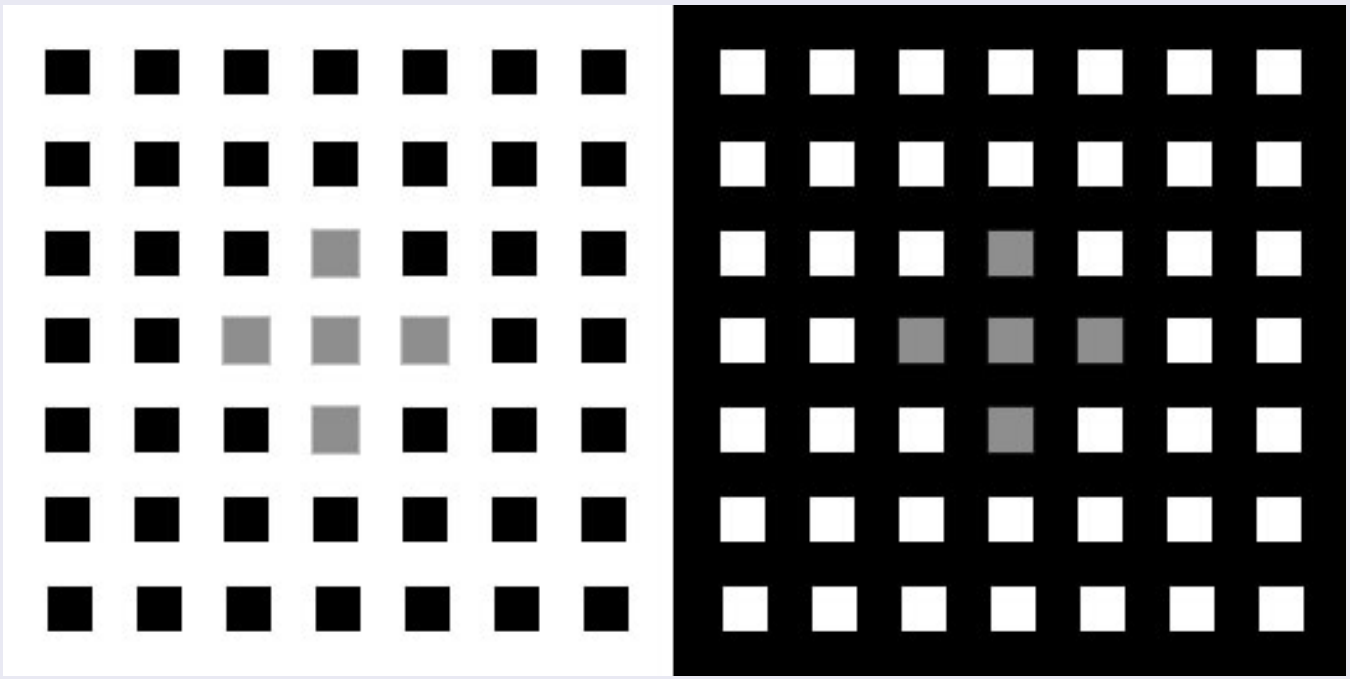

Figure: Impact of Local Context on Interpretation

The Linearisation Problem

Information Structure

- **Information Structure** describes the way we organize the contents of a sentence to help with its correct interpretation in the current context.
- Essentially, describes how we prepare a new link in the chain of sentences.

The Linearisation Problem

Theme

- An important part of the Information Structure is the **Theme** (Thematic Structure) that:
  - **Connects back** and links to the previous discourse, maintaining a coherent point of view.
  - Serves as a **point of departure** for the further development of the discourse.
- The producer can use the *Theme* to better establish the local context.

Information Structure

Theme: Example A

- "A: What is this?"
- "B: [**T** This is] nothing special"

Theme: Example B

- "A: Who owns the green car?"
- "B: John [**T** owns the green car]"

Information Structure

Theme: Example A

- "A: What is this?"
- "B: [**T** This is] nothing special"

Theme: Example B

- "A: Who owns the green car?"
- "B: John [**T** owns the green car]"

Information Structure

Theme and Rheme

Once the *Theme* has been established, the new contribution, called the Rheme can be identified. So, the Information Structure of each sentence will typically contain two parts:

- **Theme (or Topic)** (í. þema)
  Connection to previous discourse. Point of departure. What is being discussed.
  Assume Receiver knows this.
- **Rheme (or Comment)** (í? Rema)
  New contribution to the discourse. Comment about that which is discussed.
  Assume Receiver does *not* know this.

Information Structure

Theme and Rheme: Example

- "A: Where are you going tonight?"
- "B: [**T** Tonight I'm going] [**R** to the movies]"

Information Structure

Same Proposition but Different Information Structure

The same propositional content can be produced with different *Information Structure* since the required context can change.

Example

- "John hit Jack"
- "[**T** The one who hit Jack was] [**R** John]"
- "[**T** The one who got hit by John was] [**R** Jack]"

Information Structure

Using Different Discourse Devices

Different *Discourse Devices* can be used to deliver the same Information Structure. The previous example used grammatical constructs to create a different Information Structure, but you can also use intonation (the tone of the voice). This example shows this option using **boldface** to indicate a pitch accent.

Example

- "John hit Jack"
- "[**T** The one who hit Jack was] [**R** John]"
  OR "**John** hit Jack"
- "[**T** The one who got hit by John was] [**R** Jack]"
  OR "John hit **Jack**"

Information Structure

Information Structure Does Not Follow Grammatical Structure

As seen in the last example, *Information Structure* does not follow the *Grammatical Structure*. A *subject* is sometimes a *Theme*, but not always.

Example

- "Who hit Jack? [**R John**] [**T** hit Jack]"
- "Who did John hit? [**T** John hit] [**R Jack**]"

Information Structure

Reflection on The Producer/Receiver Cooperation

The Information Structure reflects what the producer believes the receiver wants or needs to hear. One can therefore say that it is a product of Grice's cooperative principle, where the maxims of Information (Quantity) and Relation (Relevance) are being balanced and honored.

(Discourse Analysis)

The Gricean Maxims

These *maxims* describe the assumptions receivers make about producers' contributions, and therefore they interpret them with this in mind:

- **Truth (Quality)**
  Do not say what you believe to be false or for which you lack evidence.
- **Information (Quantity)**
  Make your contribution informative but not too loaded.
- **Relevance (Relation)**
  Be relevant.
- **Clarity (Manner)**
  Avoid obscurity and ambiguity. Be brief and orderly.

Information Structure

Effect of Information Structure

Let's look at what we would consider a *proper* presentation of information, and then see the same propositional contents produced with different (non-cooperative) information structure (Original example is from Halliday).

News Report (cooperative)

The sun is shining, it's a perfect day. Here come the astronauts. They're just passing the Great Hall; perhaps the President will come out to greet them. No, it's the admiral who's taking the ceremony...

Information Structure

Effect of Information Structure

Let's look at what we would consider a *proper* presentation of information, and then see the same propositional contents produced with different (non-cooperative) information structure (Original example is from Halliday).

News Report (cooperative)

The sun is shining, it's a perfect day. Here come the astronauts. They're just passing the Great Hall; perhaps the President will come out to greet them. No, it's the admiral who's taking the ceremony...

Information Structure

News Report (seems un-cooperative - IS conveyed by syntax)

It is the sun that's shining, the day that's perfect. The astronauts come here. The Great Hall they're just passing; he'll perhaps come out to greet them, the President. No, it's the ceremony that the admiral's taking...

News Report (seems un-cooperative - IS conveyd by intonation)

The sun is shining, it's a perfect **day**. **Here** come the astronauts. They're just passing the **Great Hall**; perhaps the President will **come out to greet them**. No, it's the admiral who's taking **the ceremony**...

Information Structure

News Report (seems un-cooperative - IS conveyed by syntax)

It is the sun that's shining, the day that's perfect. The astronauts come here. The Great Hall they're just passing; he'll perhaps come out to greet them, the President. No, it's the ceremony that the admiral's taking...

News Report (seems un-cooperative - IS conveyd by intonation)

The sun is shining, it's a perfect **day**. **Here** come the astronauts. They're just passing the **Great Hall**; perhaps the President will **come out to greet them**. No, it's the admiral who's taking **the ceremony**...

Information Structure

Implications for NLP Systems

Taking Information Structure into account can help a machine interact naturally with a human, just like it helps humans interacting with each other.

- Understanding: A machine could extract new contributions from sentences based on the discourse devices that relate to Information Structure.
- Generation: Picking the appropriate Information Structure and producing the supporting discourse devices will help humans to understand (and appreciate) what the machine is trying to convey.

Information Structure

Implications for NLP Systems - Text-To-Speech

- Do Text-To-Speech systems model Information Structure?
- Let's try the sentence pair "Where did you get sick? I got sick on the airplane." with some Text-To-Speech (TTS) systems available on the web (next slide) and pay attention to the intonation.
- What should the correct intonation be? What is the intonation of the TTS? Does the TTS sound like it understood the question?
- The context, and therefore the Information Structure of the response, can easily be changed with different questions. For example, "What happened to you on the airplane? I got sick on the airplane."

Information Structure

Implications for NLP Systems - TTS Systems to Try

- ATT:
  http://www.research.att.com/ ttsweb/tts/demo.php
- Festival:
  http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html

Seminal Work on Using Information Structure for TTS

Hiyakumoto, L., Prevost, S., Cassell, J. (1997) "Semantic and Discourse Information for Text-to-Speech Intonation", Proceedings of ACL Workshop on Concept-to-Speech Technology

Information Structure

Implications for NLP Systems - TTS Systems to Try

- ATT:
  http://www.research.att.com/ ttsweb/tts/demo.php
- Festival:
  http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html

Seminal Work on Using Information Structure for TTS

Hiyakumoto, L., Prevost, S., Cassell, J. (1997) "Semantic and Discourse Information for Text-to-Speech Intonation", Proceedings of ACL Workshop on Concept-to-Speech Technology

Outline

Newness/Givenness of Information

Information Status (í. upplýsingaástand)

- When the producer is constructing the *Information Structure* of a new production, she needs to make assumptions about what can be taken as given in the discourse and what would constitute a new and interesting contribution.
- Prince calls this the **Assumed Familiarity** of information, but more commonly we refer to this labeling of information as **Information Status**.
- Let's assume here for simplicity that *information* refers to the contents of *Noun Phrases / Referring Expressions*.

Newness/Givenness of Information

Information Status

When a producer wishes to include a *Noun Phrase*, the *Information Status* of that noun phrase depends on the availability of a corresponding *Discourse Entity* (DE) in the *Discourse Model* (DM). The following terminology has been used to describe this status.

- **Evoked** (also "Old" or "Given")
  The corresponding DE already exists in the DM.
- **Inferrable** (also "Mediated")
  The receiver could infer the corresponding DE and add it to the DM themselves.
- **New**
  A corresponding DE does not exist and the producer explicitly creates it.

Information Status

Evoked (also Old or Given)

- If information is already in the *Discourse Model* when it is brought up, it is **Evoked** information.
  - It is **Textually Evoked** if the producer *instructed* the receiver to put it there earlier (maybe as New or Inferrable information).
  - It is **Situationally Evoked** if the receiver needed no instruction to include it in the Discourse Model (it may have been obvious from the situation, like "you" ad "me").

Example

A1: "[I] saw a brown cat, did [you] see [it]?" (Situ, Situ, Textu)

Information Status

Inferrable (also Mediated)

- Even though the information is not *Evoked*, it is possible that the receiver could infer it from the context (Cognitive, Situational or Textual).
- This may require *common sense* or *domain knowledge*

Examples

- "I can't see [the sun]" (world knowledge)
- "When I got home, [the door] was open" (part inferred from whole)
- "On the 'science trip' [the beer] was too warm" (known kind of event)

Information Status

New

- All information that is neither *Evoked* nor *Inferrable*, must be **New**.
- *New* information may need to be created from scratch before being put into the DM (Brand New) or "copied" from a corresponding entry in a different DM (Unused).

Examples

- "I saw [a blue giraffe]!" (Brand New)
- "I saw [Harrison Ford]!" (Unused)