Foundations of Natural Language Processing
Lecture 3a
Text Corpora: Motivation

Alex Lascarides

Corpora in NLP: Motivation

This lecture:

- What is a corpus?
- Why do we need text corpora for NLP? (learning, evaluation)
- Illustrative application: **sentiment analysis**

Corpora in NLP

corpus
noun, plural *corpora* or, sometimes, *corpuses*.

1. a large or complete collection of writings: the entire corpus of Old English poetry.

2. the body of a person or animal, especially when dead.

3. *Anatomy.* a body, mass, or part having a special character or function.

4. **Linguistics. a body of utterances, as words or sentences, assumed to be representative of and used for lexical, grammatical, or other linguistic analysis.**

5. a principal or capital sum, as opposed to interest or income.

Dictionary.com

Corpora in NLP

- To understand and model how language works, we need empirical evidence. Ideally, **naturally-occurring** corpora serve as realistic samples of a language.

- Aside from linguistic utterances, corpus datasets include **metadata**—side information about where the language comes from, such as author, date, topic, publication.

- Of particular interest for **core NLP**, and therefore this course, are corpora with **linguistic annotations**—where humans have read the text and marked categories or structures describing their syntax and/or meaning.

Examples of corpora (in choronological order)

Focusing on English; most released by the **Linguistic Data Consortium** (LDC):

Brown: 500 texts, 1M words in 15 genres. POS-tagged. **SemCor** subset (234K words) labelled with WordNet word senses.

WSJ: 6 years of *Wall Street Journal*; subsequently used to create Penn Treebank, PropBank, and more! Translated into Czech for the **Prague Czech-English Dependency Treebank**.

ECI: European Corpus Initiative, multilingual.

BNC: 100M words; balanced selection of written and spoken genres.

Redwoods: Treebank aligned to wide-coverage grammar; several genres.

Gigaword: 1B words of news text.

AMI: Multimedia (video, audio, synchronised transcripts).

Google Books N-grams: 5M books, 500B words (361B English).

Flickr 8K: images with NL captions

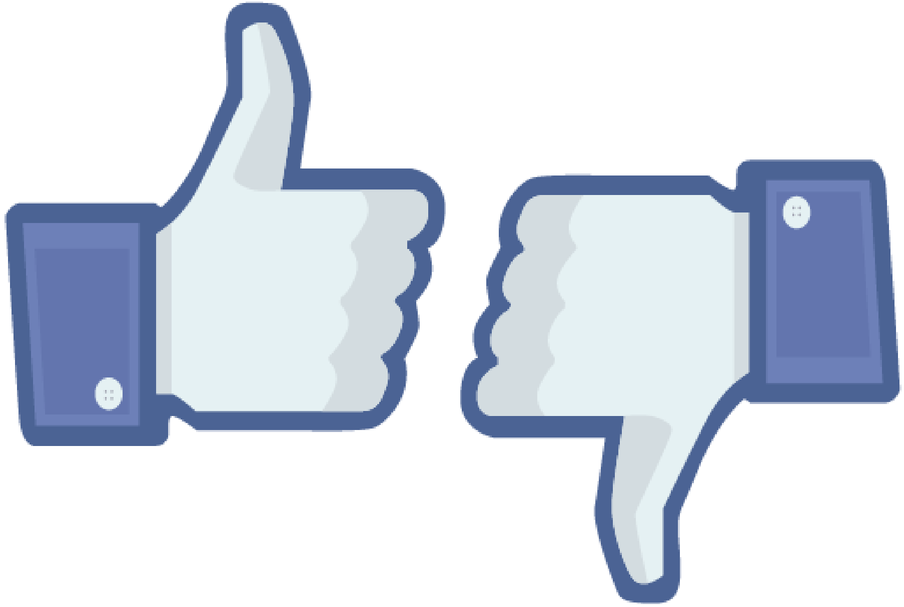
English Visual Genome: Images, bounding boxes ⇒ NL descriptions

Markup

- There are several common markup formats for structuring linguistic data, including XML, JSON, CoNLL-style (one token per line, annotations in tab-separated columns).

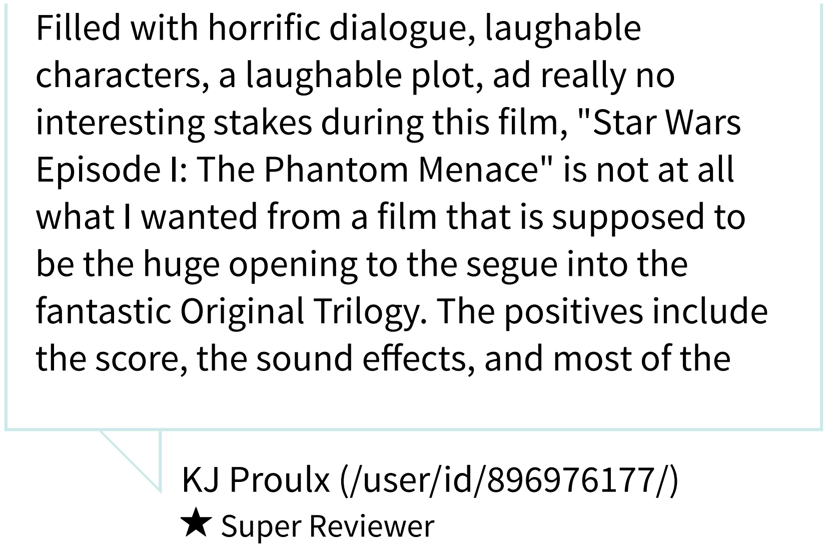
- Some datasets, such as WordNet and PropBank, use custom file formats. NLTK provides friendly Python APIs for reading many corpora so you don't have to worry about this.

Sentiment Analysis



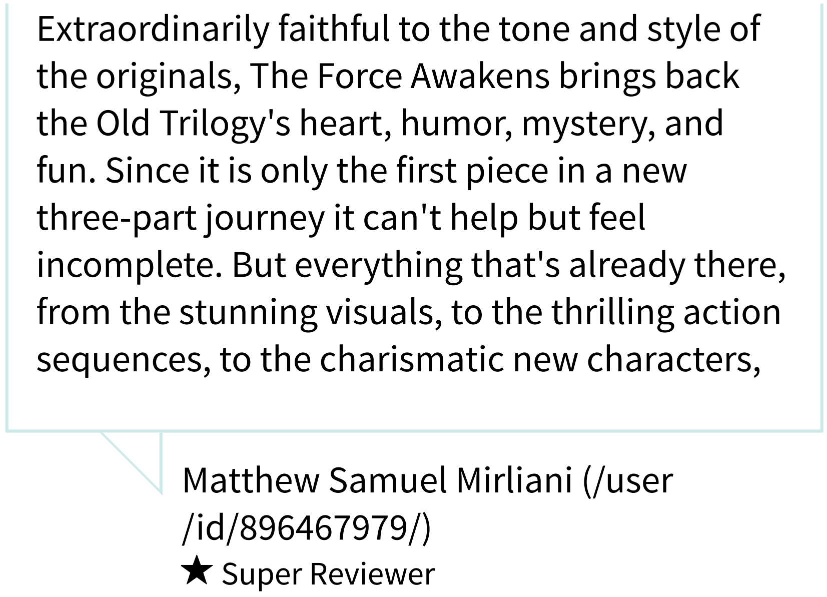
Goal: Predict the **opinion** expressed in a piece of text. E.g., + or −. (Or a rating on a scale.)

Sentiment Analysis



Filled with horrific dialogue, laughable characters, a laughable plot, ad really no interesting stakes during this film, "Star Wars Episode I: The Phantom Menace" is not at all what I wanted from a film that is supposed to be the huge opening to the segue into the fantastic Original Trilogy. The positives include the score, the sound effects, and most of the

KJ Proulx (/user/id/896976177/)
★ Super Reviewer

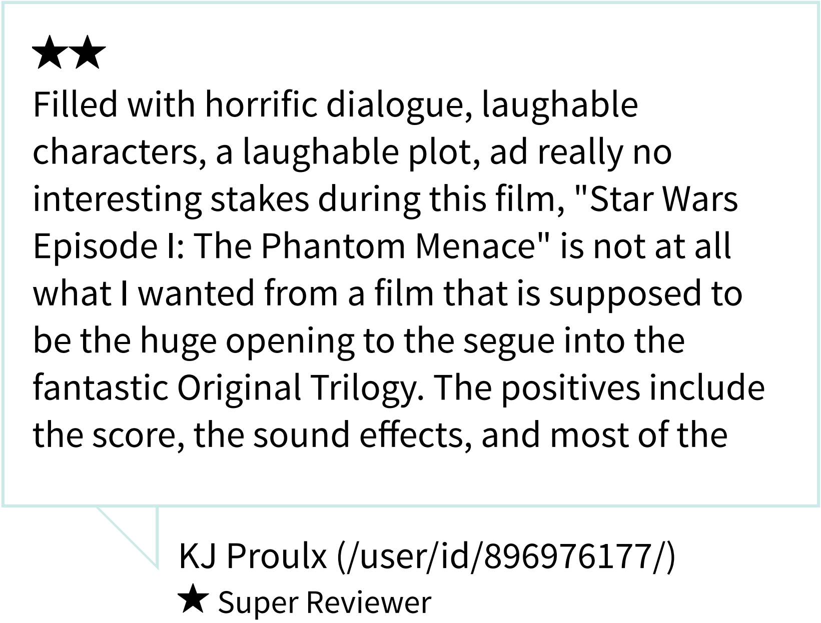


Extraordinarily faithful to the tone and style of the originals, The Force Awakens brings back the Old Trilogy's heart, humor, mystery, and fun. Since it is only the first piece in a new three-part journey it can't help but feel incomplete. But everything that's already there, from the stunning visuals, to the thrilling action sequences, to the charismatic new characters,

Matthew Samuel Mirliani (/user /id/896467979/)
★ Super Reviewer

RottenTomatoes.com

Sentiment Analysis

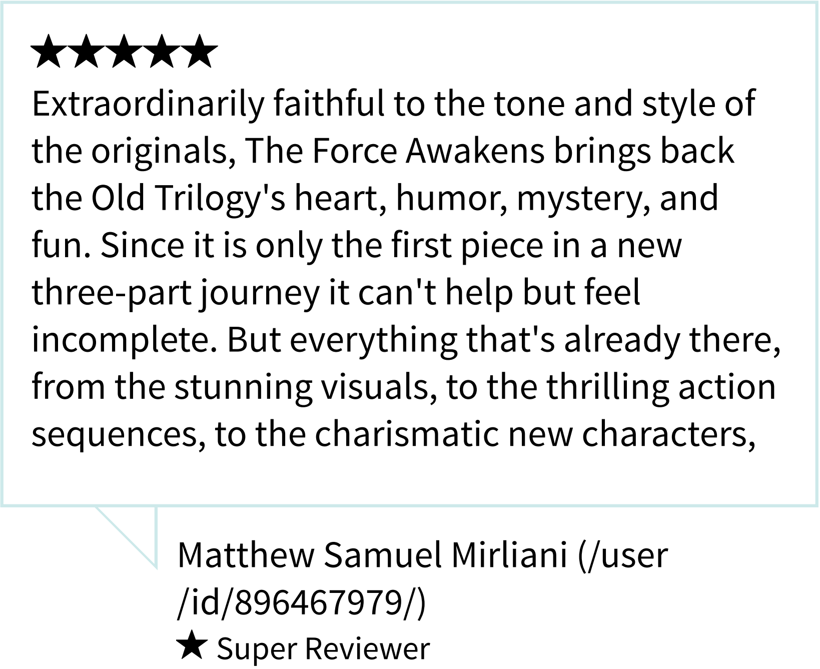


★★

Filled with horrific dialogue, laughable characters, a laughable plot, ad really no interesting stakes during this film, "Star Wars Episode I: The Phantom Menace" is not at all what I wanted from a film that is supposed to be the huge opening to the segue into the fantastic Original Trilogy. The positives include the score, the sound effects, and most of the

KJ Proulx (/user/id/896976177/)

★ Super Reviewer



★★★★★

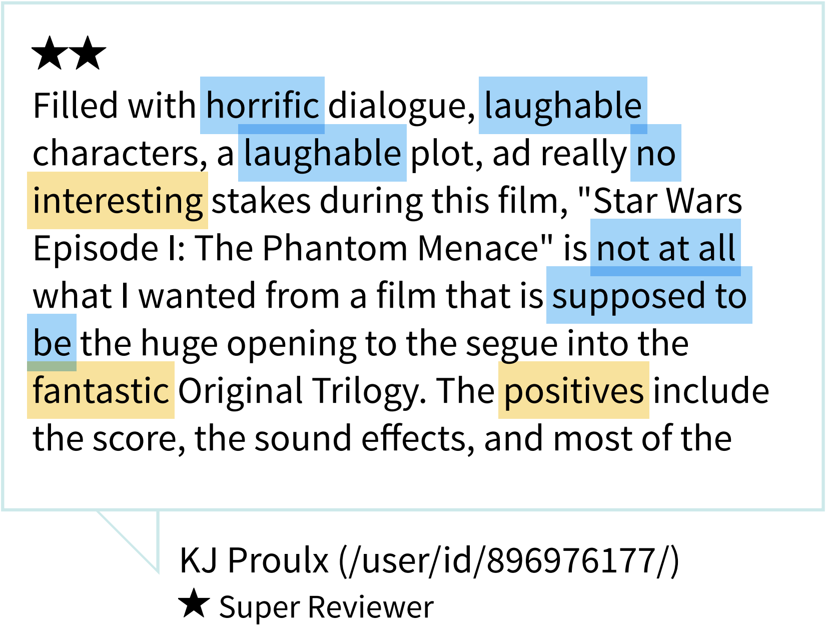
Extraordinarily faithful to the tone and style of the originals, The Force Awakens brings back the Old Trilogy's heart, humor, mystery, and fun. Since it is only the first piece in a new three-part journey it can't help but feel incomplete. But everything that's already there, from the stunning visuals, to the thrilling action sequences, to the charismatic new characters,

Matthew Samuel Mirliani (/user /id/896467979/)

★ Super Reviewer

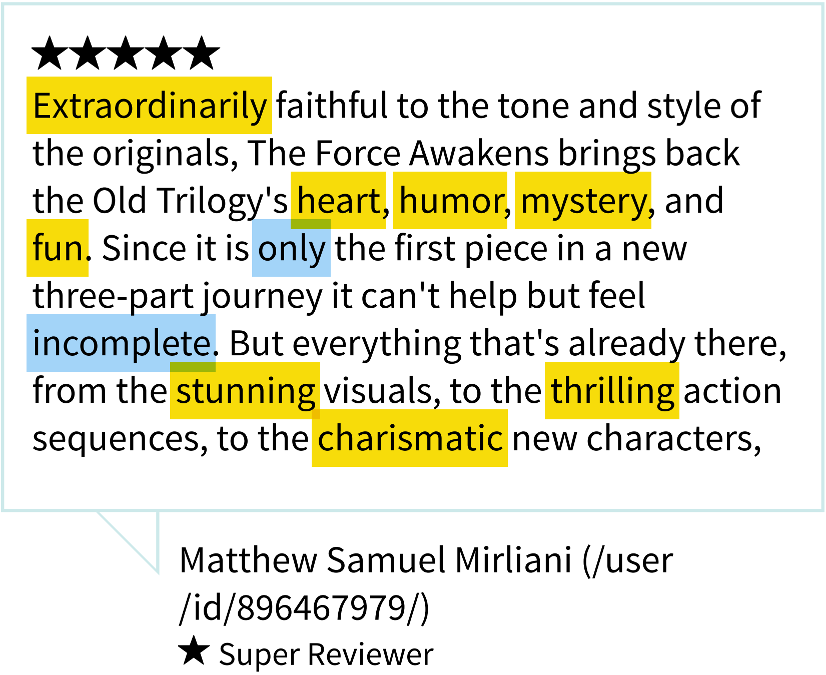
RottenTomatoes.com

Sentiment Analysis



★★
Filled with horrific dialogue, laughable characters, a laughable plot, ad really no interesting stakes during this film, "Star Wars Episode I: The Phantom Menace" is not at all what I wanted from a film that is supposed to be the huge opening to the segue into the fantastic Original Trilogy. The positives include the score, the sound effects, and most of the

KJ Proulx (/user/id/896976177/)
★ Super Reviewer



★★★★★
Extraordinarily faithful to the tone and style of the originals, The Force Awakens brings back the Old Trilogy's heart, humor, mystery, and fun. Since it is only the first piece in a new three-part journey it can't help but feel incomplete. But everything that's already there, from the stunning visuals, to the thrilling action sequences, to the charismatic new characters,

Matthew Samuel Mirliani (/user /id/896467979/)
★ Super Reviewer

RottenTomatoes.com + intuitions about positive/negative cue words

So, you want to build a sentiment analyzer

Questions to ask yourself:

1. What is the input for each prediction? (sentence? full review text? text+metadata?)

2. What are the possible outputs? ($+$ or $-$ / stars)

3. How will it decide?

4. How will you measure its effectiveness?

The last one, at least, requires data!

Summary

- Machine readable text (and speech) has revolutionised NLP

- Machine learning trained on corpora yields an empirically grounded approach to NLP applications

- All NLP tasks are highly complex, making a data-driven approach attractive.

- At the very least, you need corpora to evaluate the quality of your NLP system.

Next Time: Corpora in NLP: The basics