Corpora: they are necessary

Grammar-based approaches to opinion mining: Part 2 (ESSLLI 2013)

Asad Sayeed

Uni-Saarland

On the menu

- Corpus requirements for supervised learning.
- Examples of existing corpora.
- Crowdsourcing for sentiment analysis.

Just saying: I can't emphasize the importance of corpus construction enough.

Supervised learning: still where it's at

- News flash: opinion is *subjective*.
- We are trying to model what a person is thinking when they say something.
- We are *building systems* that are supposed to approximate a *human judgement*.

So we need to collect *data*—and maybe, lots of it!

Q: So why not just collect a lot of product reviews?

A: Well, to recap. . .

- There are more genres than just product reviews.
- Texts are complicated, contain indications of multiple opinions.
- Complex relationships between words, sentences, and so on within text have an effect on interpreting opinions.
- Complex relationship with the world. . .

But we can't handle the entire world.

So how to construct a data source?

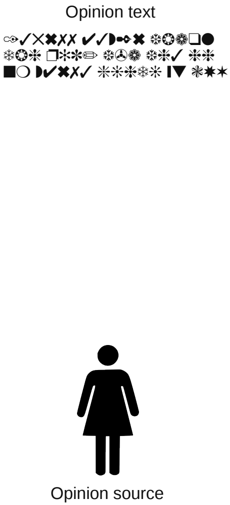
1. Task dependence – need to understand what information is really required to model opinion.
2. What exists already:
   - Other resources may have solved a related problem – bootstrap?
   - Existing automatic tools may solve part of the problem reliably (e.g., POS tagging).
3. Selecting/finding/training annotators.
4. Validating the annotation.

Presenting the challenges may be better diagrammatically

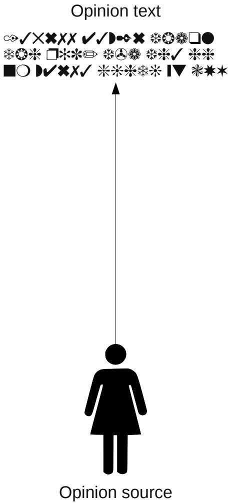
Start by considering the opinion source

Opinion source

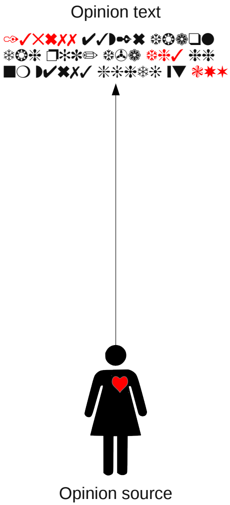
and some opinion text



generated by the source.



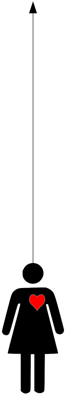
Parts of the text indicate intensity, polarity, etc.



But consider a reader

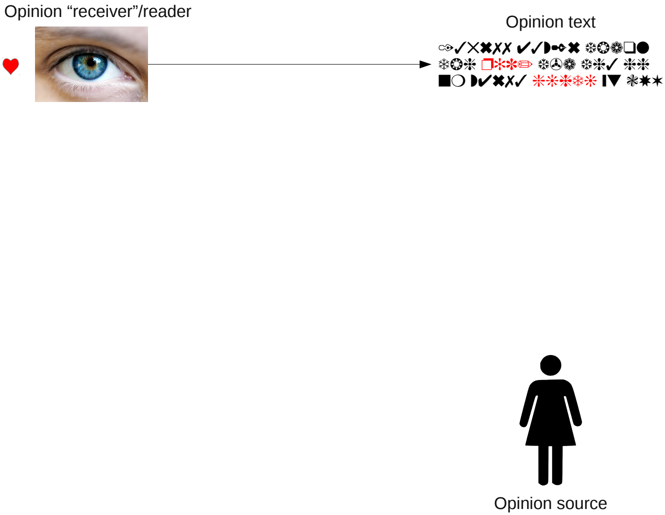
Opinion “receiver”/reader

Opinion text

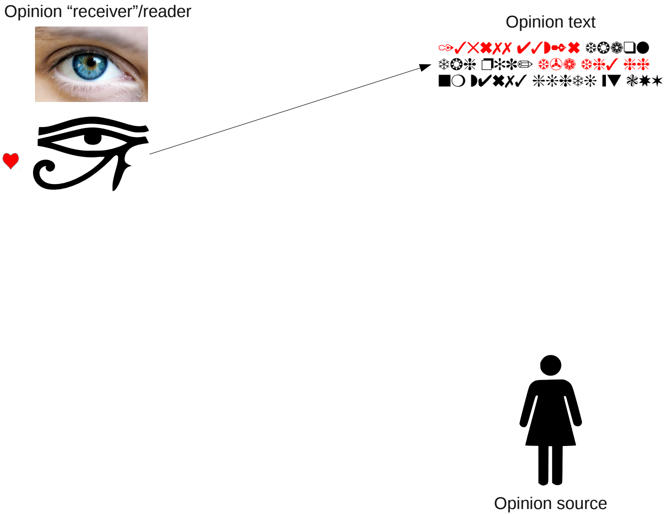


Opinion source

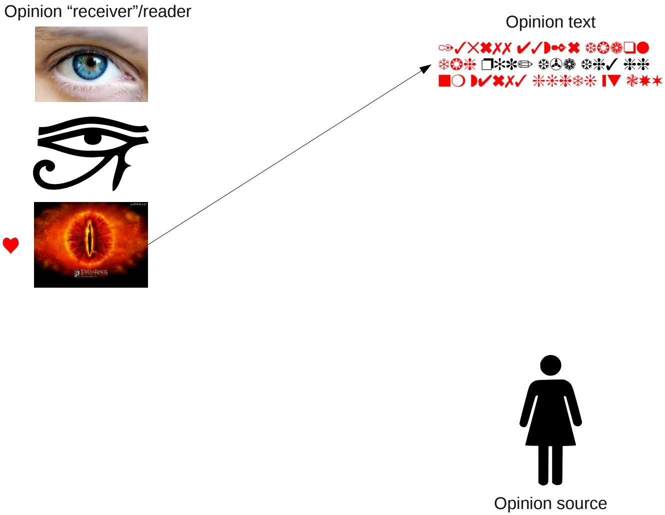
whose understanding of the opinion in the text is different.



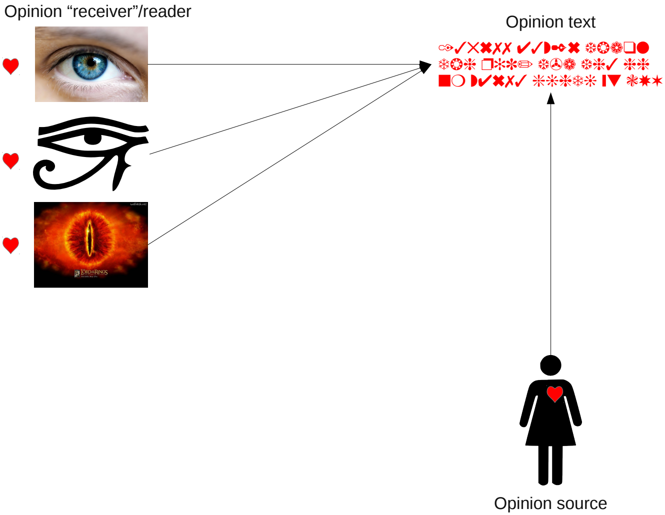
whose understanding of the opinion in the text is different.



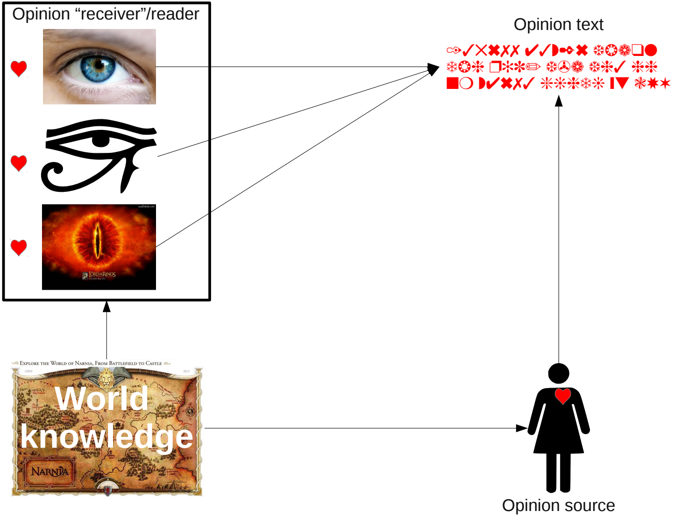
whose understanding of the opinion in the text is different.



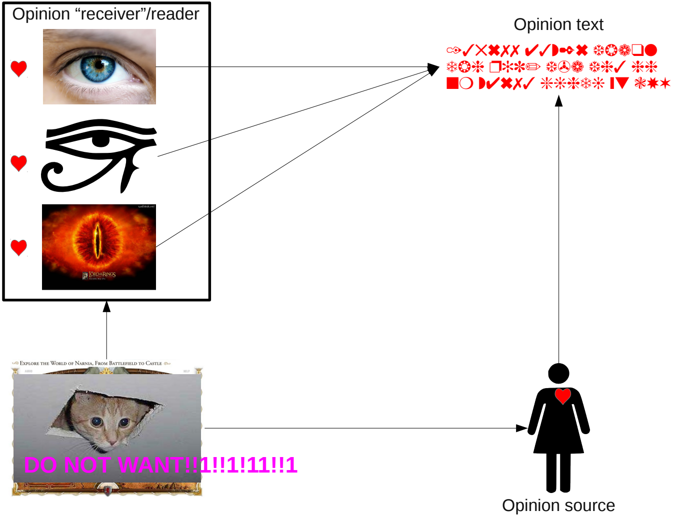
Thus it becomes hard to identify the important bits.



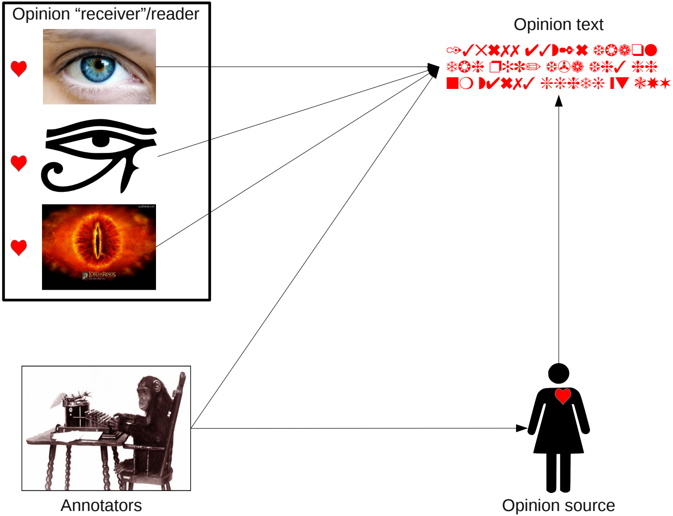
Disentangling them takes pragmatics.



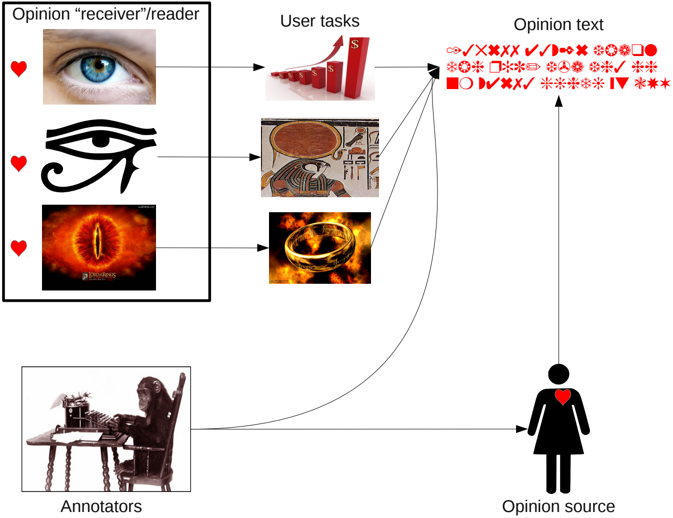
But pragmatics is hard.



Annotators contain some world knowledge,



but are they annotating something useful for opinion mining users?



An example application: the IT business press.

As we discussed yesterday:

- Want to predict the evolution of an opinion community – in order to predict the evolution of technological innovation.
- Focused on particular relevant concepts (opinion targets).
- Attempt to use media (IT business press) as evidence base for theorizing.

What does the task look like?

Example: information technology business press

Lloyd Hession, chief security officer at BT Radianz in New York, said that virtualization also opens up a slew of potential network access control issues.

- Choose a {*source, target, opinion*} triple: {Lloyd Hession, virtualization, negative}
- What is the evidence that Mr. Hession has a negative opinion about virtualization?

Deciding on the source is easy for people.

Example: information technology business press

Lloyd Hession, chief security officer at BT Radianz in New York, said that virtualization also opens up a slew of potential network access control issues.

- This can be done at high recall/precision through grammatical clues (Choi et al., 2006).

Deciding on the target is not so easy

- But finding the target is not so easy:

Example: information technology business press

Lloyd Hession, chief security officer at BT Radianz in New York, said that virtualization also opens up a slew of potential network access control issues.

- What is the evidence that it should be "virtualization" and not e.g. "access"? Or both?
- From whose perspective does it matter?
  - Mr. Hession's personal opinion might be positive.
  - In the information technology press (IT), the market reaction might be more important.

So back to the pragmatic factors again.

Example: information technology business press

> *Lloyd Hession, chief security officer at BT Radianz in New York, said that virtualization also opens up a slew of potential network access control issues.*

- But if we're modeling a business journal, then what matters is the *reader*—the market.
- Either way, need world-knowledge: "pragmatic opinion" (again, Somasundaran and Wiebe, 2009).

But say we "fix" source and target ID

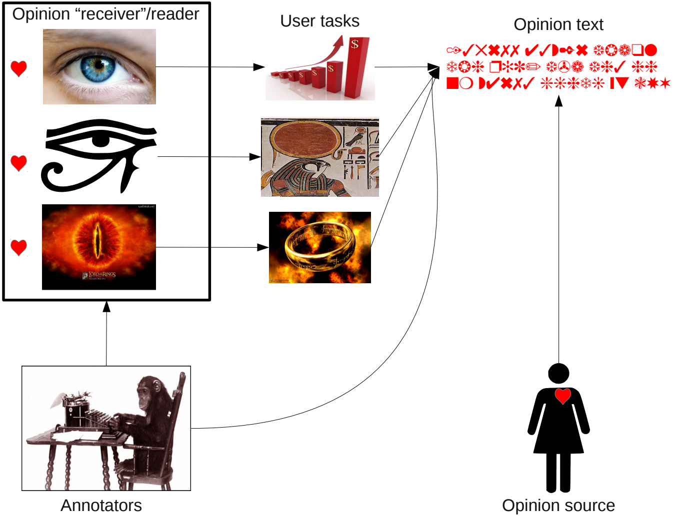
Example: information technology business press

Lloyd Hession, chief security officer at BT Radianz in New York, said that virtualization also opens up a slew of potential network access control issues.

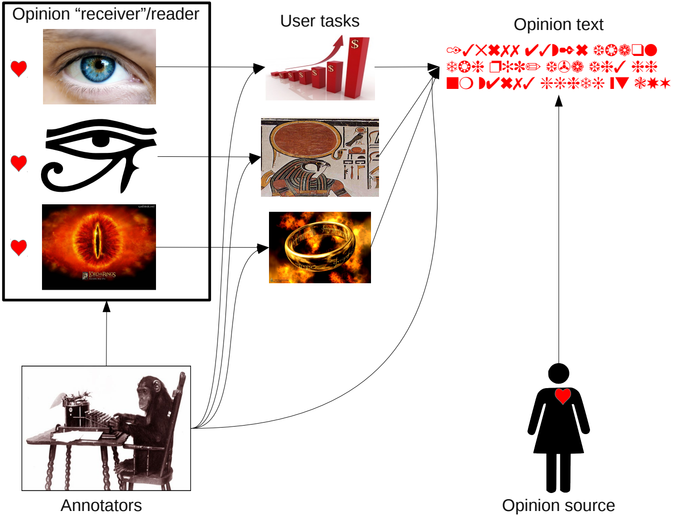
- "slew" and "issues": convey negative sentiment about "virtualization".
- How do we know they're negative in this domain?
- What about words like "update"? Important in IT domain, not mentioned in major polarity lexicon.

The "little" details of syntax/semantics and the "big" details of pragmatics actually intertwine.

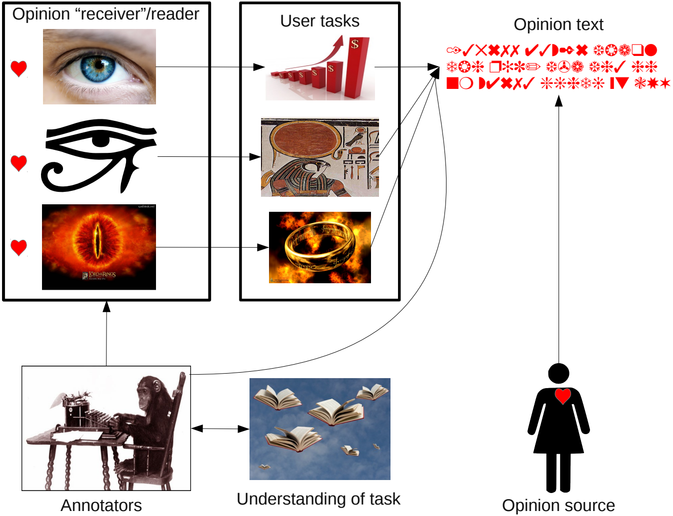
So we need to find a way to connect annotation to the users,



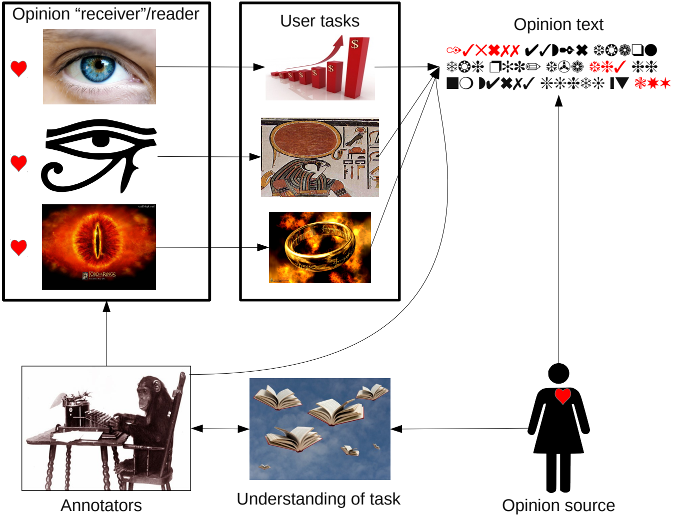
and that involves understanding the tasks.



But no two users are alike in knowledge and understanding,



although they seem to match opinion sources relatively well.



How do you evaluate that, anyway?

A very common way is via Cohen's κ.

Measures agreement between two annotators given discrete categories for a given set of objects.

- Invented originally for psychology.
- One of a number of measures, but most common.
- Better than simple percent agreement as (in theory) it takes into account agreement accruing to chance.

Despite weaknesses, it's practically *de rigeur* to report it, or an alternate statistic.

How do we calculate Cohen's κ

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

- $\Pr(a)$ = percent observed agreement among raters.
- $\Pr(b)$ = probability of random agreement.
  - For a given label s and raters r_1 and r_1, calculate $\Pr(s|r_1)$ and $\Pr(s|r_2)$.
  - Multiply. That's how likely they were to agree on on that class.
  - Sum over all classes s. That's how likely they were to agree on all classes.

How do we interpret Cohen's κ?

No one knows!

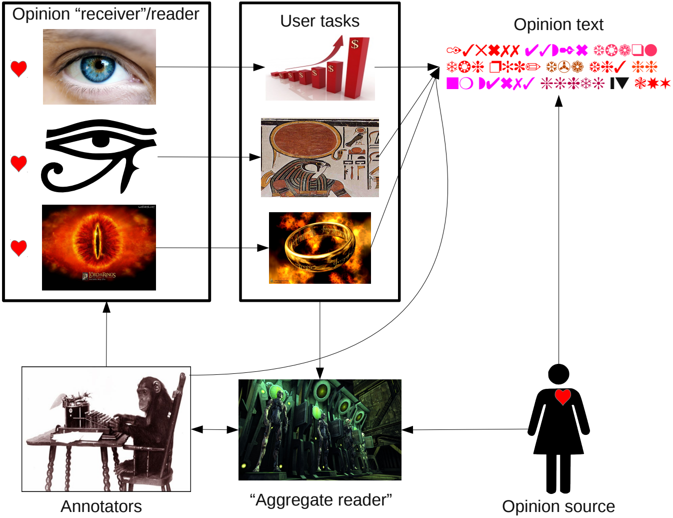
- The Landis and Koch "guidelines": 0-0.20 is "slight", 0.21-0.40 is "fair", 0.41-0.60 is "moderate", 0.61-0.80 is "substantial", 0.81-1 is "perfect".
- Negative values possible! (It's not a true probability.)
- Statistical significance – difficult to calculate.
- Tend to report results at "substantial" agreement.
- Difficult to compare across tasks – what might be low in one case may be high in another, who knows?

So are there alternatives?

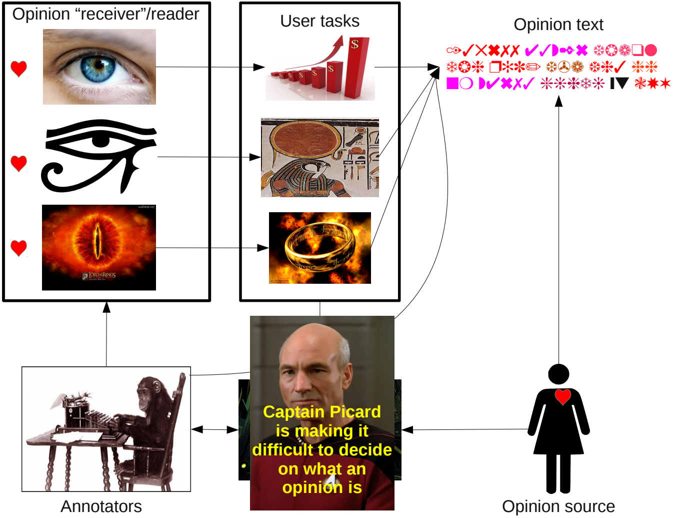
Yes.

- Scott's pi – two annotators.
  - Calculates $\Pr(e)$ differently: assumes response distribution same for both annotators.
  - Possibly weaker than Cohen's κ (for that reason)
- However, Scott's "pi" can be extended to Fleiss' κ – allows multiple annotators.
- Just plain old precision and recall against a gold standard.
- Bayesian approaches (Carpenter 2008).

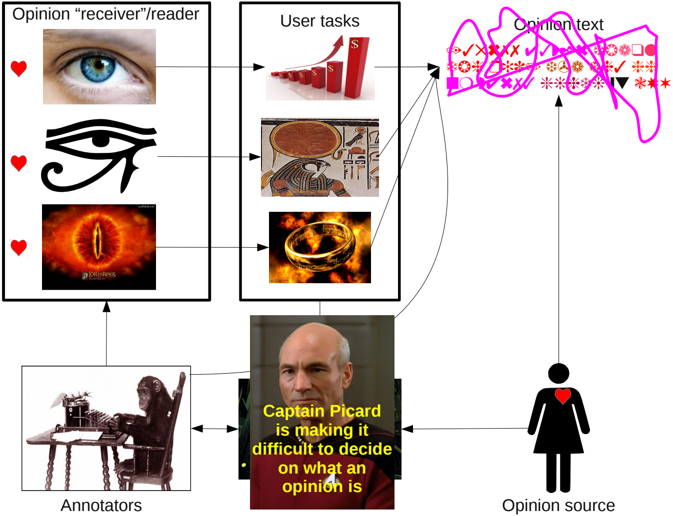
How to bring task understanding into the mix?



Ensuring multiple annotators agree on what an opinion is



is crucial to consistently identifying opinion-relevant text.



So what resources are there for fine-grained sentiment analysis, anyway?

An example: the MPQA

Multi-Perspective Question-Answering corpus (Wiebe et al. 2009).

- Central goal: annotate "private states".
  - Hidden (unverifiable) variable inside an individual consciousness.
  - Can be feelings, goals, opinions, etc.
  - Subjective content (as opposed to fact) as evidence for private state in text.
- Annotate data in text that has multiple perspectives with different latent "private states".
  - ie, not single-author reviews—opinions entangled with one another.

How are private states expressed?

- Explicit mentions of private states:

‘

‘The US **fears** a spill-over,” said Xirao-Nima.

- Speech events expressing private states:

‘

‘The report is full of absurdities,” Xirao-Nima **said**.

- Expressive subjective elements:

‘

‘The report is **full of absurdities**,” Xirao-Nima said.

Explicit mentions and speech events = “direct subjective frames”.

And how are these annotated in the MPQA?

The *span* of the body text of an MPQA subjective frame is called a “text anchor”. Then they annotate:

- Source: person/entity expressing evidence for private state.
- Target: topic of frame/what the opinion is about (only for direct subjective).
- Attitude type (polarity) and intensity of private state.
- Expression intensity - contribution of expression to intensity. (only direct subjective)
- Insubstantial – true/false e.g. if referring to a hypothetical (only direct subjective).

Many other annotation details we won’t get into, and more recent updates.

What does the MPQA contain?

As of Wiebe et al. (2005):

- 10,657 sentences in 535 documents from 187 news sources from June 2001 to May 2002.
- Three annotators, two students and one library archivist, 8-12 hours per week, 3-6 months each.
- Agreement measured by text anchor overlap.
  - So a match between two annotators can have some words missing.
  - They use symmetric pairwise agreement rather than κ: $\text{agr}(A \| B) = |A \text{matches} B| / |A|$
- Implications for very fine-grained/grammar-based systems?

Let's take a look at the MPQA texts directly for a moment.

It's not the only thing around.

The J.D. Power and Associates (JDPA) Corpus (Kessler et al. 2010).

- A "sentiment corpus for the automotive domain."
- It annotates:
  - Entity mentions and coreference (as well as meronymy).
  - Sentiment expressions – with their links to entities and polarity.
  - Modifiers:
    - Sentiment negators.
    - Neutralizers – expressions that do not commit speaker to truth of sentiment expression.
    - Committers – shift certainty of sentiment.
    - Intensifiers.

What does it look like?

A sentiment expression

My friends and family feel extremely safe in our Hummer.

A couple of committers

A good-looking car in itself . . .
The interior looks to be in a nice condition.

What's in the JDPA and how was it made?

- 335 blog posts consisting of 233,001 tokens.
- Annotators: there were 7.
    - Trained via pilot project after reading guidelines – correction and feedback.
    - Iterative: 8 batches, which guideline changes between batches as necessary.
    - Not all annotators marked all documents!

Calculating agreement: complicated.

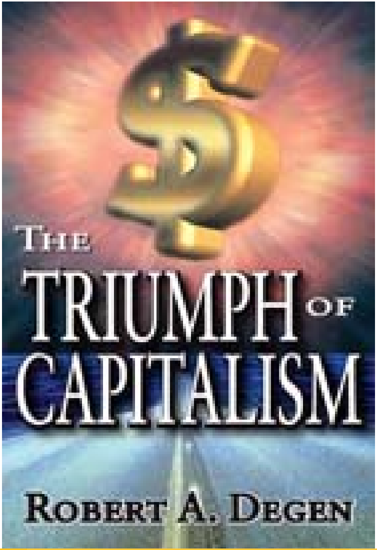
Annotation	Property	Type	Agreement	Matched
Mention	—	span	83%	21,518
Mention	Semantic Type	property	83%	17,923
Mention	MentionPriorPolarity	property	100%	7
Mention	ContextualSentiment	property	95%	13
Mention	EntitySentiment[1]	property	85%	87
Mention	Inferred Contextual Sentiment[2]	property	87%	18,706
Mention	Refers-to	span-entity-link	68%	5,684
Mention	Part-of	entity-entity-link	35%	1,178
Mention	Feature-of	entity-entity-link	23%	294
Mention	Member-of	entity-entity-link	81%	34
Mention	Instance-of	entity-entity-link	73%	184
SentimentExpression	—	span	75%	3,976
SentimentExpression	PriorPolarity	property	95%	3,712
SentimentExpression	Target	span-entity-link	66%	2,879
Negator	—	span	66%	384
Negator	NegatorTarget	span-span-link	85%	335
Neutralizer	—	span	36%	70
Neutralizer	NeutralizerTarget	span-span-link	78%	64
Intensifier	—	span	60%	729
Intensifier	IntensifierDirection	property	96%	690
Intensifier	IntensifierTarget	span-span-link	95%	737
Committer	—	span	33%	93
Committer	CommitterDirection	property	91%	79
Committer	CommitterTarget	span-span-link	82%	75
OPO	—	span	33%	93
OPO	OPOTarget	span-span-link	66%	383

But resource creation is expensive.

Annotators insist on eating.

Fortunately, we have a way of fending off hungry annotators!

It's called crowdsourcing: the ultimate commodification of labour.



And here's where it all began...

Crowd + outsourcing = crowdsourcing

- Outsourcing
  - Came into common use in the 90s.
  - Practice of cutting costs by getting work done by outside firms.
  - Specifically came to refer to "labor arbitrage", particularly "offshoring"
    - (Indian and Chinese workers cheaper, don't go on strike...)

. . . and how it works.

Basic idea

- Small "clerical" and mechanical tasks can already be performed cheaply. (ie, outsourcing, offshoring)
- Intellectual tasks: not necessarily that complicated!
  - A lot of "complex" tasks can just be broken down.
- Get multiple people to do (repeatedly) the same small task unit.
- Need some exterior method to validate and consolidate results.
  - e.g. Inter-annotator agreement, statistical aggregation techniques.

And here's where you should use it.

"Asad's law": that I just made up

Any intellectual task can be crowdsourced if and only if you can turn it into a series of yes/no questions.
(This is not investment advice.)

And here's where you should use it.

"Asad's law": that I just made up

Any intellectual task can be crowdsourced if and only if you can turn it into a series of yes/no questions.
(This is not investment advice.)

. . . BUT . . .

And here's where you should use it.

"Asad's law": that I just made up

Any intellectual task can be crowdsourced if and only if you can turn it into a series of yes/no questions.
(This is not investment advice.)

...BUT ...

That doesn't mean you SHOULD turn it into a series of yes/no questions.

It's mostly done on Mechanical Turk.

Need an intermediary.

- Online payment is a huge problem—most projects do not have the resources/expertise to manage this!
- Legal issues (incl. international), tax issues, dealing with banks and credit card companies.
- Even big companies balk.

Biggest crowdsourcing platform: Amazon Mechanical Turk

- If there's anyone who knows about payments, it's Amazon.
- What is Turkish about the Mechanical Turk? It's a long story.

So who are the workers anyway?

Extensively studied by Prof. Panos Ipeirotis (at New York University). Some sample statistics/fun facts (Ipeirotis 2010):

- Overwhelmingly from USA (46%) and India (34%).
- Americans overwhelmingly female (almost 70%) and Indians overwhelmingly male (almost 70%).
- Population surprisingly young and educated (both US and India).
- US household income not particularly high, but a small number are surprisingly wealthy.
  - Beware, though, of statistical aggregation hiding more complex social patterns...

Now that we know something about the population. . .

. . . we can speculate about why it costs so very little.

- Most jobs posted on Amazon Mechanical Turk pay (US) pennies per HIT ("Human Intelligence Task").
- We will see some in the demo, and you may understand better why.
- (But consider that each one must be done multiple times.)

Something that particularly applies to sentiment/opinion.

User interface design is **Critical**.

- Using UNSKILLED labour (whose familiarity with computers may be more limited than you think).
- Using anonymous labour
  - NEWSFLASH: There are dishonest people on the Internet.
  - Consider how easy it is to answer a yes/no questionnaire by only clicking "yes".
  - MTurk has a way to block users...but is that too coarse-grained a quality control?
- **Especially for comp ling**: the tasks people want to do via MTurk are often quite technical!
- Some help from third-party platforms built on top of MTurk.

Some interesting NLP examples

Groningen Meaning Bank

- Closest thing I could find to a treebank.
- Semantically annotated corpus.
- Crowdsourced annotation via a sort of game, plus a wiki for expert annotators.

Machine translation

- Bloodgood and Callison-Burch (2010): Urdu-English translation at 10 cents a sentence (rather than 10 cents a word for other non-crowdsources translation efforts.)

Q: How fine-grained an annotation can crowdsourcing produce?

A: Even down to the word level.

Tootin' my horn again: Sayeed et al. (2011).

- How to generate a resource that has annotations down to the word level?
- How to evaluate it on strict boundaries?
- How to get random untrained people on the internet to perform a complex, domain-dependent task?

What resources were used?

- Information technology (IT) business journal: InformationWeek
  - Approx 33K articles of varying lengths (news blurbs, full-feature articles).
  - Approx 75K sentences containing IT concept mentions.
    - OpenNLP splitter.
  - Years covered: 1991-2008.
- IT concepts: same 60 including "Enterprise resource planning", "application service provider", and abbreviations and variants.

What was needed to make the interface work?

- Interface desiderata:
    - It should be no easier to cheat than to answer intelligently.
    - It should be more fun than a multiple-choice test.
    - It should make the “what is an opinion” decision implicit in the task.

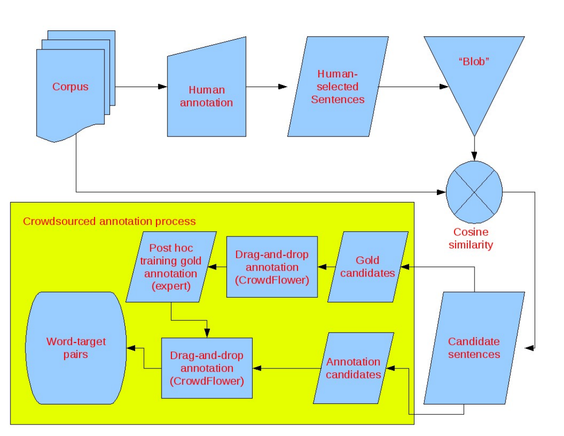
What was needed to make the interface work?

- Drag and drop interface task design on CrowdFlower:
  - One sentence per task, plus context.
  - Candidate words highlighted.
  - Four boxes: positive, negative, no opinion, can't tell.
  - Each highlighted word must be dragged to one box.
- Boxes are equally "difficult" to drag words to—not much harder to answer intelligently than to cheat.
- "What is an opinion" is not directly asked.

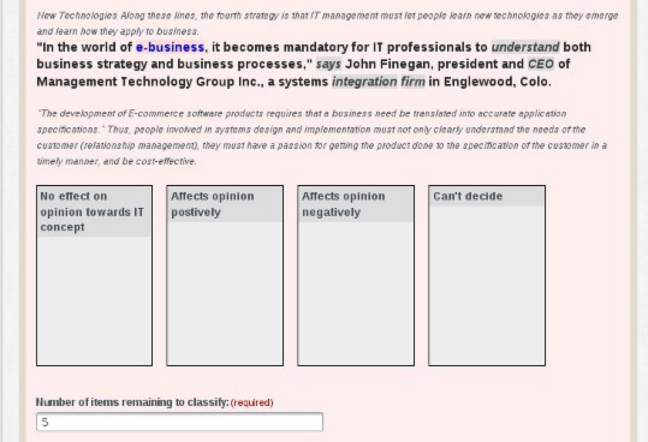
Everything looks better with flowchart.

- Schematic view:

Drag-and-drop interface strategy



What did it look like?


New Technologies Along these lines, the fourth strategy is that IT management must let people learn new technologies as they emerge and learn how they apply to business.

"In the world of e-business, it becomes mandatory for IT professionals to *understand* both business strategy and business processes," *says* John Finegan, president and *CEO* of Management Technology Group Inc., a systems *integration firm* in Englewood, Colo.

"The development of E-commerce software products requires that a business need be translated into accurate application specifications." Thus, people involved in systems design and implementation must not only clearly understand the needs of the customer (relationship management), they must have a passion for getting the product done to the specification of the customer in a timely manner, and be cost-effective.

No effect on opinion towards IT concept	Affects opinion postively	Affects opinion negatively	Can't decide

Number of items remaining to classify: (required)

5

Now select data from the corpus.

- IW corpus divided into sentences.
  - Filtered for direct mentions of IT concepts.
- Sentences selected by human annotators aggregated into a single string and converted to unit vector.
- Cosine similarity
  - Every sentence scored against the string, ranked.
  - Top N represent "first-pass" selection of sentences likely to contain opinion words related to targets.

Give the workers a lighter load.

- Need to select candidate words in the sentences.
  - But we don't know in advance what the opinionated words are likely to be in this domain.
- Use Stanford POS tagger to select open-class words.
- For each sentence, randomly group selected words into "highlight groups" of six.
  - Non-overlapping groups.
  - Reduce the number of classifications required from workers.

Word highlight group

The amount of industry attention paid to this new class of integration software speaks volumes about the need to extend the reach of **ERP** systems.

Give the workers a lighter load.

- Need to select candidate words in the sentences.
  - But we don't know in advance what the opinionated words are likely to be in this domain.
- Use Stanford POS tagger to select open-class words.
- For each sentence, randomly group selected words into "highlight groups" of six.
  - Non-overlapping groups.
  - Reduce the number of classifications required from workers.

Word highlight group

The amount of industry attention paid to this new class of integration software speaks volumes about the need to extend the reach of **ERP** systems.

How to evaluate the output?

- CrowdFlower—automated quality control above MTurk.
- Basis of quality control: training "gold".
  1. Posted 50 highlight groups on CrowdFlower with no training gold (3 users/task at 3 cents/task).
  2. *Post hoc* "correction" of result by us. This is the training gold data.
  3. Our gold contained only "obvious" cases.
- Gold items are included randomly in tasks. Workers rejected if below 65% correctness.
- CrowdFlower's handling of correctness has limitations (either "all correct" or "one correct", no "minimum correctness").

And then we ran it.

- 200 highlight groups. (approx 1200 highlighted words)
- 3 users/task, 4 cents/task. (Total \$60, incl. CrowdFlower fees.)
- Aggregation: majority vote, with ties going to "no opinion".
- Task finished overnight.
- CrowdFlower's task apportionment means that some tasks had 4+ answers.

But we're not done yet.

- Baseline: assign, wherever possible, the polarity from the Pitt sentiment lexicon, or none if unavailable.
- Stringent filtering
  - CrowdFlower's quality control still pretty generous to workers.
    - With good reason: too hard-nosed will reject workers too quickly.
  - Score every worker by strict compliance with gold (accuracy).
  - Remove bottom *n* workers. (Some units may be lost.)
- Evaluation: on 30 tasks done by us.
  - Retrieval task: precision, recall
  - Agreement as Cohen's κ—not often used in fine-grained sentiment analysis.

And we evaluated it.

Excluded	**Words lost (of 48)**	**Prec/Rec/F**	**Cohen's κ**
(prior polarity)	N/A	0.87 / 0.38 / 0.53	*-0.26*
0	0	0.64 / 0.71 / 0.67	0.48
1	0	0.64 / 0.71 / 0.67	0.48
3	0	0.66 / 0.73 / 0.69	0.51
5	0	0.69 / 0.73 / 0.71	0.53
7	2	0.81 / 0.76 / **0.79**	**0.65**
10	9	**0.85** / 0.74 / **0.79**	0.54
12	11	0.68 / 0.68 / 0.68	0.20

Now that we've talked about resources, we need to talk about doing something with them.

These days, we tend to subject corpora to machine learning. . .