Opinion mining: the larger context

Grammar-based approaches to opinion mining: Part 1 (ESSLLI 2013)

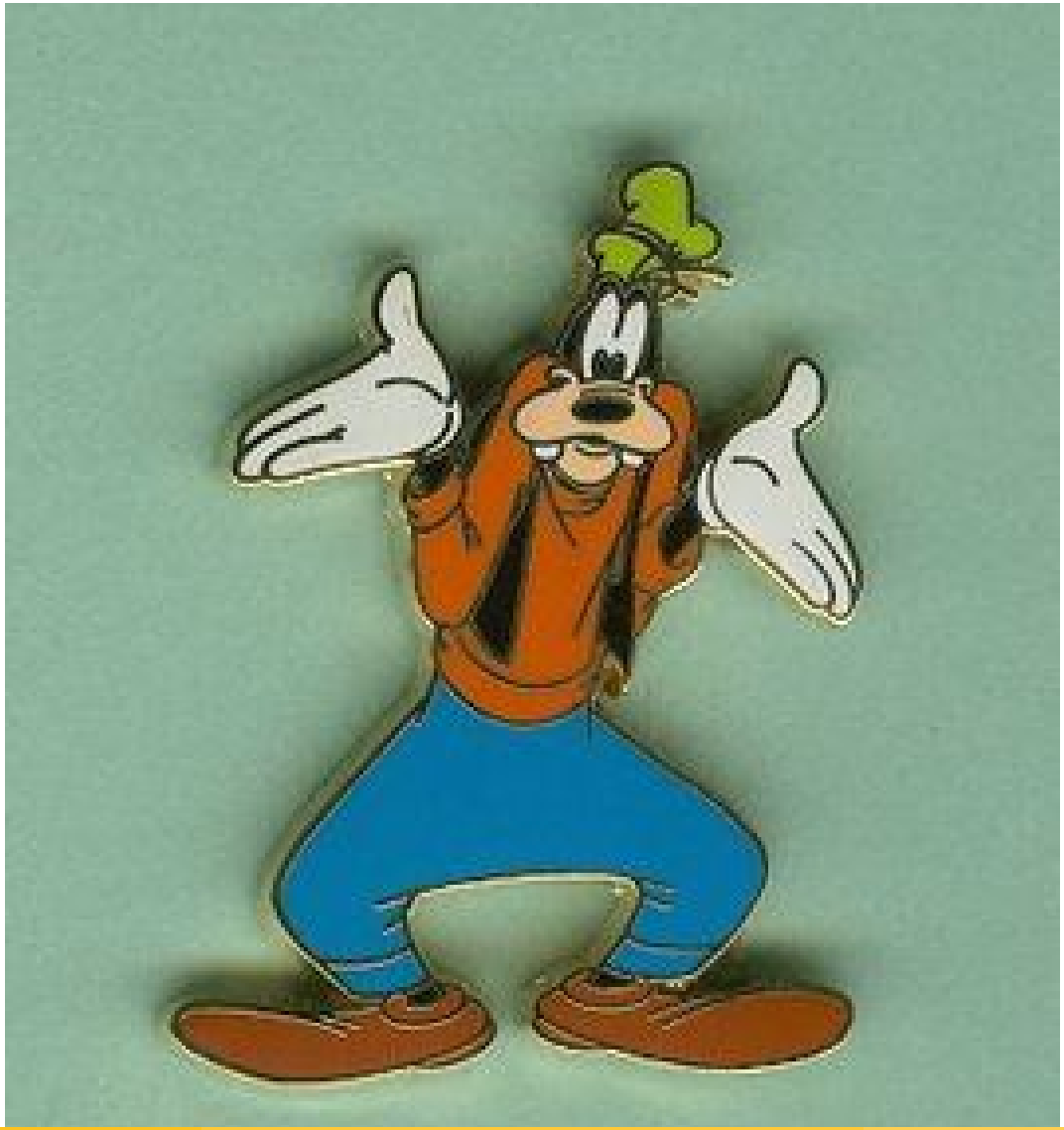
Asad Sayeed

Uni-Saarland

Q: What is sentiment analysis?

(Often used interchangeably with "Opinion Mining".)

A: Nobody knows!



Despite that, sentiment analysis is everywhere.

For the time being. . .

Non-definition of opinion mining

. . . let's just "define" it operationally in terms of opinion mining:
An *opinion source* holds an *opinion/sentiment* about an *opinion target*.

Of course, that begs the question.

Q: What is an opinion?

And we can go on from there.

But we won't.

Q: What does “grammar-based” mean here?

A: Overall, dependent on some finer-grained syntactic (and/or semantic) property of the language being opinion-mined.

What this course will do.

- Provide a high-level overview of opinion mining practices.
- Make a case for the use of grammar in opinion mining.
- Focus on the use of grammatical structures to identify opinionated language.
- Describe recent techniques and technologies – and by "recent" I mean from now to 10-12 years ago, though usually not more than 7-8.

Also. . .

- My perspective: won't pretend to be unbiased. I have an opinion about opinions.
- A dialogue – feel free to raise anything you like, anytime.

Some of the things we'll cover

- Part 1: Context and foundations
  - Sentiment analysis background: product reviews, debates.
  - Challenges of perspective, pragmatics.
  - Corpus-based social science.
- Part 2: Corpus resources
  - Desiderata for fine-grained sentiment analysis corpora.
  - Existing examples of corpora (MPQA, JDPA).
  - Crowdsourcing.

Some of the things we'll cover

- Part 3: Machine learning refresher – "bird's eye" view.
- Part 4: Sequence-based techniques
  - Source and target identification.
  - HMM/CRF-based techniques.
- Part 5: Structure-based techniques
  - Unsupervised structure discovery.
  - Machine learning over structures.

However, it's a really big field: we can only do a sample in five sessions!

Q: So, uh, why grammar?

A: Depends on what you're doing

A simple case: movie/product reviews

2 of 2 people found the following review helpful

★★★★★ **Amazing... LOVE it!!**, September 19, 2012

By **E. Share** - See all my reviews

REAL NAME

This review is from: Barefoot Running - The Movie: How to Run Light and Free by Getting in Touch with the Earth (NTSC/US Version) (DVD)

Wow! I just finished watching this and it is amazing. I'm completely blown away by the stunning cinematography. The movie was almost entirely filmed in Maui and I didn't want to take my eyes off the gorgeous colors on the screen. The beaches, trails, mountains and roads that Michael and Jessie run on throughout the movie, are absolutely breathtaking. There is even a bonus feature section that displays Michael's photography in a zen-like production of photographs that he captured during his runs.

The movie is designed as easy to follow, clear, concise chapters. It takes the viewer from the first step of shedding your shoes and putting your bare feet to the earth to running like a child again with a light foot and a free spirit. I am not a runner, yet this movie encourages me to take the journey of my first few yards without shoes. I plan to watch it over and over again while I practice what I've learned. This is an amazing resource guide put to life!

I think this movie is a terrific instructional video for experienced athletes too, who want to halt running injuries (that are discussed in detail in the movie) and find strength and speed that they probably never had.

This movie is filled with information such as (i) discussing "how to" go barefoot, (ii) what kind of shoe to purchase for the times when you must have a shoe, (iii) exercises to help your body heal before and between runs, and (iv) even ways of bringing the earth's energy into your body to help with overall health. It documents a serious accident that Michael had, which was the catalyst in getting him into barefoot running in the first place. Michael's story is incredibly inspirational and to see the strength of his muscles up close through the filming, after viewing the footage of what his body had been through prior to going barefoot, is fascinating.

The goal: to predict the rating from the text

(Fairly safe) assumptions

- Text and rating produced by same person.
- Text and rating reflect same opinion. (Is this really safe?)
- Evidence for the rating appears in the text.
- Text contains opinionated/affective/emotional language.

(Q: Why is rating prediction a goal?)

(A: To deduce why people like what they like.)

(A2: To sell things to them. . . of course.)

Q: So how do you use text to predict what people will like?

A: Where everything starts: bag-of-words approaches.

Let's start with Turney 2002

"Thumbs up or thumbs down? Semantic orientation appled to unsupervised classification of reviews"

Process:

1. Identify phrases containing adjectives and adverbs via POS tagging.
2. Estimate the "semantic orientation" of identified phrases (pos/neg).
3. Assign "recommended" / "not recommended" labels to reviews based on average semantic orientation of phrases.

Q: From where do we get semantic orientations?

A: From Pointwise Mutual Information (PMI)

PMI

$$\text{PMI}(word_1, word_2) = \log_2 \left(\frac{p(word_1 \& word_2)}{p(word_1)p(word_2)} \right)$$

- $p(word_1 \& word_2)$ – probability that $word_1$ and $word_2$ co-occur.
- $p(word_1)p(word_2)$ – probability that they co-occur *assuming independence*.
- Probabilities estimated from search engine (PMI-IR).
- **What it means intuitively**: how well one word predicts the presence of the other, above what you'd expect if they were independent.

Then add the "semantic" sauce.

Semantic orientation of a phrase

$$\mathrm{SO}(\textit{phrase}) = \mathrm{PMI}(\textit{phrase}, \text{“excellent”}) - \mathrm{PMI}(\textit{phrase}, \text{“poor”})$$

- Another way of saying: how well the given phrase is associated with an assumed positive word vs. an assumed negative word.
- Choice of "excellent" vs. "poor" – the labels on the scales for star ratings.
- More negative $\rightarrow$ stronger association with poor?

And it's (surprisingly?) quite good!

Table 5. The accuracy of the classification and the correlation of the semantic orientation with the star rating.

Domain of Review	Accuracy	Correlation
Automobiles	84.00 %	0.4618
Honda Accord	83.78 %	0.2721
Volkswagen Jetta	84.21 %	0.6299
Banks	80.00 %	0.6167
Bank of America	78.33 %	0.6423
Washington Mutual	81.67 %	0.5896
Movies	65.83 %	0.3608
The Matrix	66.67 %	0.3811
Pearl Harbor	65.00 %	0.2907
Travel Destinations	70.53 %	0.4155
Cancun	64.41 %	0.4194
Puerto Vallarta	80.56 %	0.1447
All	74.39 %	0.5174

Q: Can we think of any pitfall??

A: Some pitfalls I can think of

(Keep in mind that this paper was very early—2002, an eon ago.)

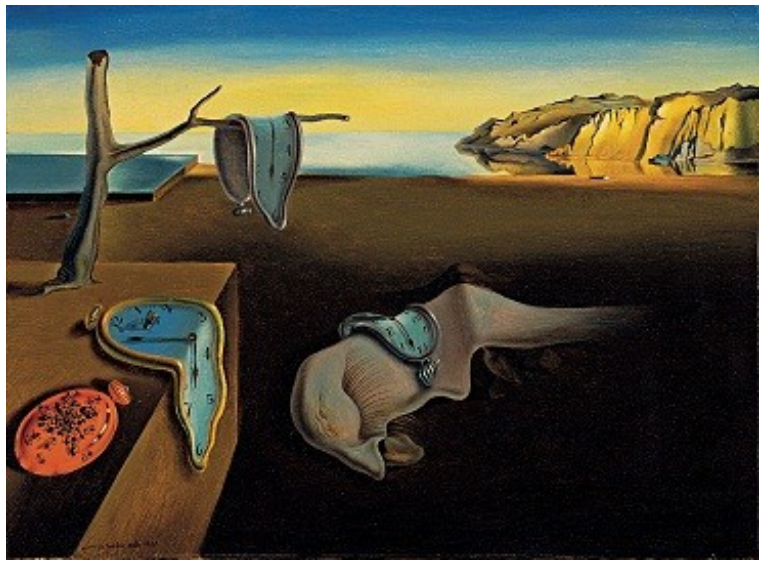
- Assumptions about collocations – high PMI with "excellent" means positive.
  - "I rarely ever find mobile carrier policies to be excellent." (made-up example)
  - But the results speak for themselves, don't they? Or is 80% enough?
- Unidimensional analysis of sentiment – quite a common problem.

OK, so we can predict recommendations from review text. So what?

The next step: what *about* a review text produced that recommendation

- What piece of language? Turney 2002/PMI might tell us that. Maybe.
- But is that enough?
- What about the product did the reviewer write positive/negative things?

The answer is “aspects” or “product features”



Possible “product” features of a surrealist painting

- Colourfulness
- Imaginativeness
- Number of melting clocks
- Famousness of artist

Widely cited: Popescu and Etzioni 2005 (EMNLP)

OPINE system:

- Input: product and corresponding reviews.
- Output: Product features and associated opinions
  - Associated opinions ranked basked on strength ("abominable" worse than "bad."

Goal Given product class C with instances I and reviews R, OPINE's goal is to find a set of (feature, opinions) tuples $\{(f, o_i, ...o_j)\}$ s.t. $f \in F$ and $o_i, ...o_j \in O$, where:

a) F is the set of product class features in R.

b) O is the set of opinion phrases in R.

c) f is a feature of a particular product instance.

d) o is an opinion about f in a particular sentence.

d) the opinions associated with each feature f are ranked based on their strength.

Pointwise Mutual Information again!

This time, PMI is used to extract feature-relevant phrases.

1. Extract noun phrases from reviews.
2. Find web-based PMI scores between NPs and "meronymy discriminators" for product class.
   - e.g., for "Scanners", meronymy discriminators include "of scanner", "scanner comes with".
3. PMI scores are used in Naive Bayes classifier, to decide which product features are relevant.

Q: And what are the product features used for?

A: Extracting opinion phrases.

- Uses dependency parser (MINIPAR) to find opinion phrase heads related to product feature phrase.
- Hard-coded dependency templates to extract phrases.

Rule Templates	Rules
$dep(w, w')$	$m(w, w')$
$\exists v$ s.t. $dep(w, v), dep(v, w')$	$\exists v$ s.t. $m(w, v), o(v, w')$
$\exists v$ s.t. $dep(w, v), dep(w', v)$	$\exists v$ s.t. $m(w, v), o(w', v)$

Table 5: **Dependency Rule Templates For Finding Words w, w' with Related SO Labels** . OPINE instantiates these templates in order to obtain extraction rules. Notation: dep=dependent, m=modifier, o=object, v,w,w'=words.

- ("Relaxation labelling" classifier for opinion phrase strength.)
  - (So we won't go into the details of how opinion strengths are found).

How well does the prod. feature extraction work?

Feature extraction relative to a previous experiment (Hu and Liu, 2004):

Data	Explicit Feature Extraction: Precision				
	Hu	Hu+A/R	Hu+A/R+W	OP/R	OPINE
D_1	*0.75*	+0.05	+0.17	+0.07	**+0.19**
D_2	*0.71*	+0.03	+0.19	+0.08	**+0.22**
D_3	*0.72*	+0.03	+0.25	+0.09	**+0.23**
D_4	*0.69*	+0.06	+0.22	+0.08	**+0.25**
D_5	*0.74*	+0.08	+0.19	+0.04	**+0.21**
Avg	*0.72*	+0.06	+ 0.20	+0.07	**+0.22**

Table 2: **Precision Comparison on the Explicit Feature-Extraction Task.** OPINE's precision is 22% better than Hu's precision; Web PMI statistics are responsible for 2/3 of the precision increase. All results are reported with respect to Hu's.

Data	Explicit Feature Extraction: Recall				
	Hu	Hu+A/R	Hu+A/R+W	OP/R	OPINE
D_1	*0.82*	-0.16	-0.08	-0.14	**-0.02**
D_2	*0.79*	-0.17	-0.09	-0.13	**-0.06**
D_3	*0.76*	-0.12	-0.08	-0.15	**-0.03**
D_4	*0.82*	-0.19	-0.04	-0.17	**-0.03**
D_5	*0.80*	-0.16	-0.06	-0.12	**-0.02**
Avg	*0.80*	-0.16	-0.07	-0.14	**-0.03**

Table 3: **Recall Comparison on the Explicit Feature-Extraction Task.** OPINE's recall is 3% lower than the recall of Hu's original system (precision level = 0.8). All results are reported with respect to Hu's.

(A refresher on precision and recall)

Sometimes you need more detail than accuracy in information retrieval.

- How do you distinguish between *error because something wanted was not found* and *wrong because something found was not wanted*?
- If true positives = tp, true negatives = tn, false positives = fp and false negatives = fn, then:
  - **Precision**: tp / (tp + fp) – how many of the things you found did you really want to find?
  - **Recall**: tp / (tp + fn) – how many of the things you wanted to find, you found?

Some lessons

- When all we want is to classify the reviews, bag-of-words is pretty good (Turney, 2002).
- But then, product features:
  - When we drill down to even a little bit of detail, we start seeing the use of grammar.
  - Not always so in other NLP tasks! Takes longer to "get to grammar".
  - But at this level hand-coded is good enough...

. . . or is it good enough?

For both Turney (2002) and Popescu and Etzioni (2007):

- We tend to get performance scores in the 80% range.
- For all I know, that might be OK for some kinds of trend analysis.
- But any more interesting "downstream" processing is going to be affected by the missing 20%.

And what might be interesting "downstream"? And what is in that missing 20%?

Let's divert our attention to debate.

Somasundaran and Wiebe (2009):

- "Debate-side" classification: figuring out who is on which side of a given debate.
- e.g. Which mobile phone is better, iPhone or BlackBerry?
  - (Remember this is 2009. And in any case the answer was and is clear: BlackBerry.)
  - (Another reminder: the question isn't "more popular"...)
- An iPhone fan may argue that the iPhone is better and/or that BlackBerry is worse
  - Need to recognize what opinion statements are *about*: the opinion target.

What makes debate special?

Somasundaran and Wiebe list some of the particular aspects of the debate genre:

- Multiple polarities to argue for a side – ie, can use positive and negative arguments to make a point.
- Sentiments towards both sides in a single contribution.
- Differentiating aspects and personal preferences – they evaluate aspects/features of the target.
- Concessions.

We see some of the things coming up from Turney (2002) and P&E (2007) here too.

An example of their text

Windows vs. Mac debate

Apples are nice computers with an exceptional interface. Vista will close the gap on the interface some but Apple still has the prettiest, most pleasing interface and most likely will for the next several years.

- On whose side is this? What tells you that?
- Observe the different ways opinion is expressed here. "...will close the gap..."

So how do they classify?

They first need to find the opinion phrase-target relations.

1. Look up words in subjectivity lexicon – there are a few of these incl. Wilson et al. (2005)
   - 8000 opinion-bearing words with positive (+), negative (-), and neutral (*) polarity.
2. Rule-based system over Stanford dependency parser output (e.g. "target is direct object of opinion word".
3. Now we have word-target pair. To eliminate sparseness, convert word to polarity. ("pleasing interface" $\rightarrow$ "interface+")

Also need to identify target aspects

Product features again!

- Web data to the rescue again!
- Downloaded avg. 3000 documents per debate using target as Yahoo search keywords (ie, "iPhone").
- Find all polarity words, and find their targets as in the previous slide – those targets are product features.
- Then calculate conditional probability that a particular feature with a particular polarity predicts a opinion towards a particular "side".
  - e.g., what is the probability that interface+ predicts Mac+?

Actually doing the classifying?

So now that Somasundaran and Wiebe have a way to

1. Identify words with opinion polarity.
2. Identify aspects of topics to which they belong.
3. Identify how aspect-opinion pairs predict stance towards the topic.

how to put this information into a model that predicts the stance from a forum contribution?

It's an optimization problem.

They find two scores:

$$w_j = P(topic_1^+ | target_i^p) + P(topic_2^- | target_i^p)$$
$$u_j = P(topic_1^- | target_i^p) + P(topic_2^+ | target_i^p)$$

- w_j and u_j correspond to the two debate sides, and j represents the target-word/polarity pair instance.
- Then these scores represent how likely it is that a particular target word favours one side or another.
- They use Integer Linear Programming to maximize the sum of all w and u.
- The side that maximizes the sum represents the class of the debate contribution.

So how well does it work?

Depends on the baseline you measure it against.

- Tested on debates from convinceme.net.
- Compared against baselines based on PMI (OpPMI) and sentiment towards topic (OpTopic) word only (ie, only opinions that specifically mention 'iPhone').

A sample result

	OpTopic	OpPMI	OpPr	OpPr + Disc
Firefox Vs Internet explorer (62 posts)				
Acc	33.87	53.23	64.52	66.13
Prec	67.74	60.0	64.52	66.13
Rec	33.87	53.23	64.52	66.13
F1	45.16	56.41	64.52	66.13

Their system is OpPr, with a version that includes discourse info.

Error analysis is where the rubber hits the road.

So where do they say their system goes wrong?

- False lexicon hits – they rely on an opinion lexicon, but sometimes an opinion word has non-opinionated senses.
- Opinion-target pairing – the rule-based syntactic system they use misidentifies these.
- Pragmatic opinion – sometimes you need world-knowledge.

> The blackberry is something like $150 and the iPhone is $500. I don't think it's worth it. You could buy a iPod separate and have a boatload of extra money left over.

The bigger picture, again.

The "leaky roof" problem – like everything else in computational linguistics.

- The more ground you cover, the more you need to cover.
  - What some people refer to as "AI-completeness".
  - To do more detailed sentiment tasks, you need more detailed grammar, more world-knowledge.
  - And these come with their own further knowledge requirements.
- This is true of most of comp ling/NLP – but requirements differ for different tasks. (machine translation vs. opinion mining?)

So let's talk world knowledge.

Since we've already talked about grammar.

- Pretty much everything about sentiment is dependent on a huge amount of "hidden" knowledge.
- PMI, use of dependency relations, and so on is *masking* or covering for our inability to keep a handle on it.
  - Unless you subscribe to some really strong form of the "distributional hypothesis."
- What technologies we can implement depend on how well we can overcome this.

What constitutes opinion-relevant world-knowledge?

The blackberry is something like $150 and the iPhone is $500. I don't think it's worth it. You could buy a iPod separate and have a boatload of extra money left over.

Evidence for the influence of pragmatics:

- Price difference – but note that $500 can be a *selling point*.
- Able to buy iPod separate (sic) – therefore not "worth it".
- Having a "boatload" of money left over is good!

But maybe, we can still find helpful evidence in the text.

The blackberry is something like $150 and the iPhone is $500. I don't think it's worth it. You could buy a iPod separate and have a boatload of extra money left over.

Possibly dispositive evidence:

- The polarity of the price difference comment is disambiguated by the next statement
  - "I don't think it's worth it" – lower price, better.
  - But even that is world knowledge...
- "Buying an iPod separate" is held to be a good thing – how would we know this.

If we keep going, the question of sentiment starts getting increasingly congruent to our understanding of the social world.

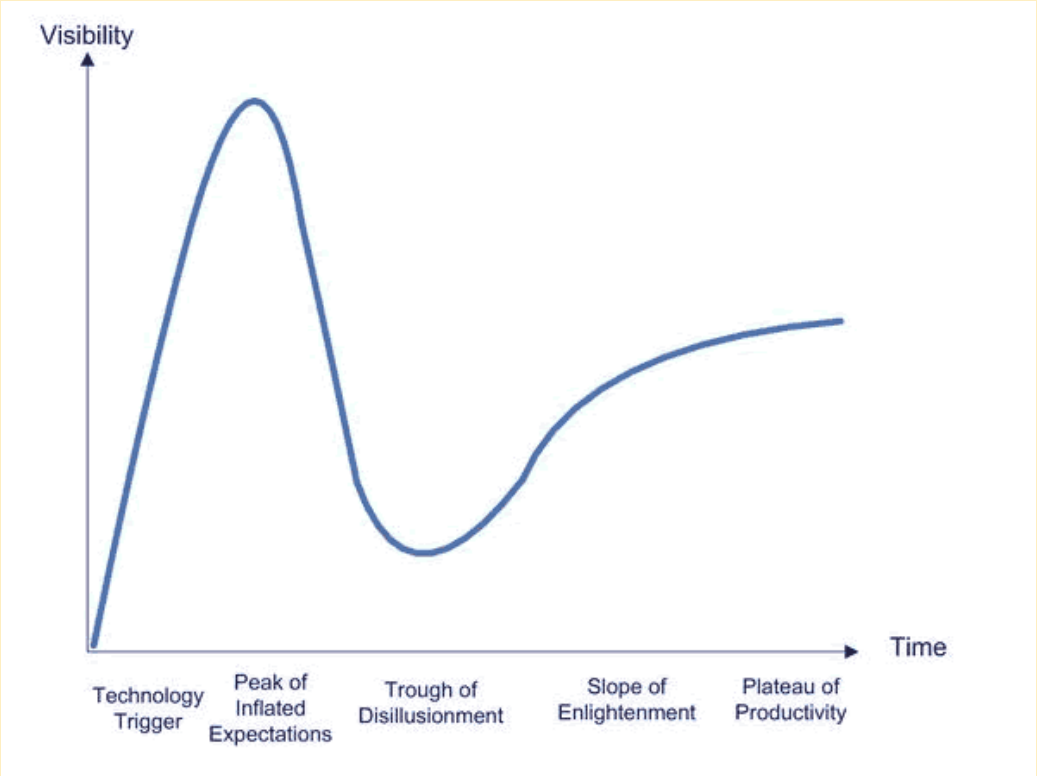
Corpus-based social science

Disclaimer: IANASS (I Am Not A Social Scientist.) But some things are clear:

- Increasingly relevant question: how do people form opinions in the first place?
- What is the environment in which opinions propagate? (Media, education, and so on.)
- How do we measure this environment?
- (. . . How do we affect this environment? . . .)

Innovation: one example of a domain where this is relevant

The Gartner Group's "Hype Cycle": just one theory of technology propagation over time.



Another way of looking at it

Tsui et al. (2009) "Understanding innovations through computational analysis of discourse" (I'm a coauthor).

- "Innovation concepts" compete with each other as alternative solutions.
- Compete for the attention of people and organizations in communities.
- Interrelated with one another in a network or "ecological system", like species.
- Communities of interest emerge and form *discourse*—what has been written and said about the innovation.

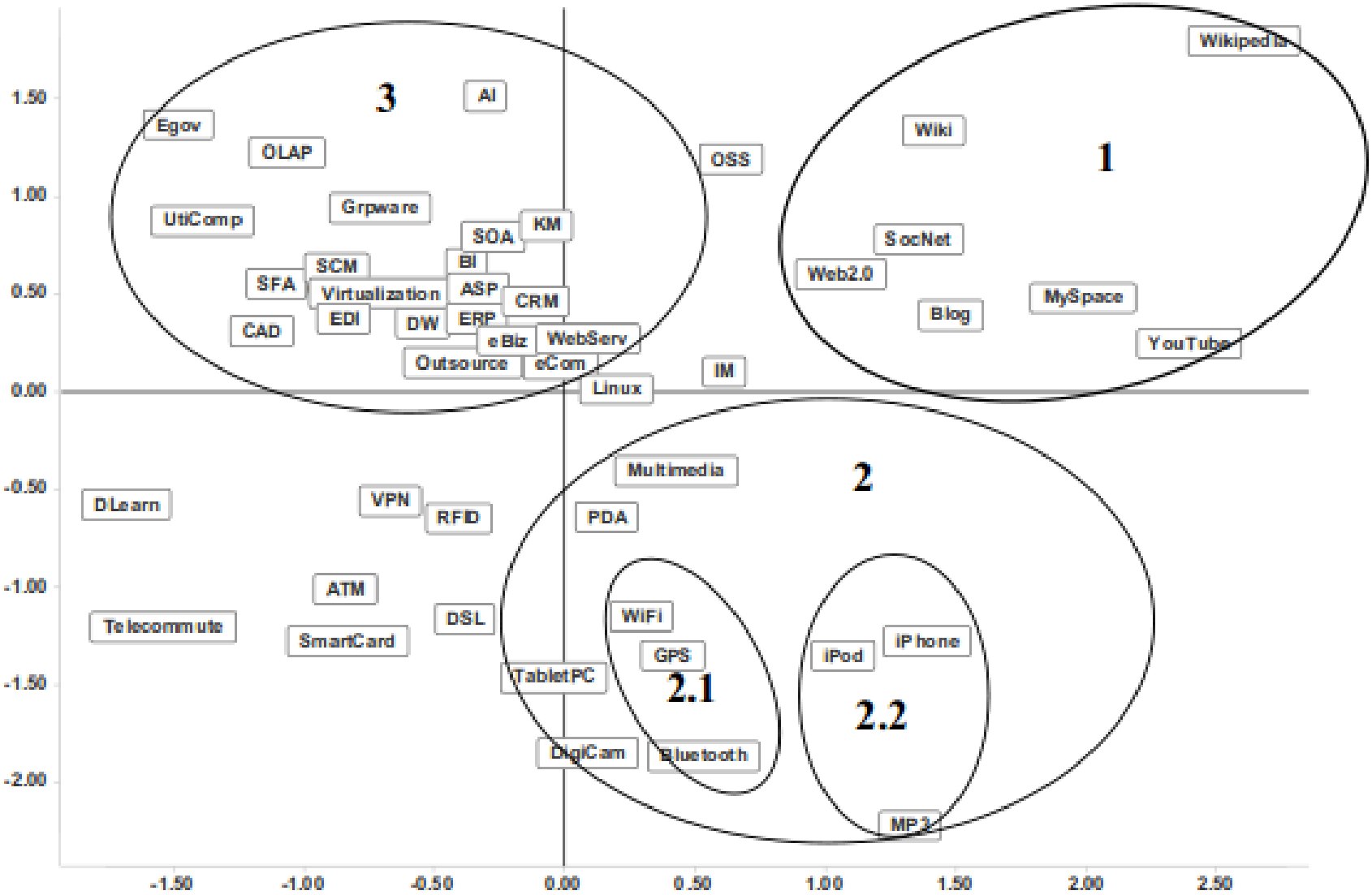
Q: Can we exploit discourse data to understand the emergence of technologies in the world?

A: Yes.

This is large scale *corpus-based social science* – attempting to validate hypotheses about our social world.

- But the state of it is very crude.
  - We took thousands of articles from IT business journals.
  - We just did a keyword search on them for paragraphs that contain members of a list of IT concepts (e.g. "WiFi", "outsourcing").
  - Calculated the divergence (KL) in the distribution of words in paragraphs per term.
  - We get n x n (where n is the number of search terms) measures of distance between concepts.

We used a clustering technique (multi-dimensional scaling).



Doesn't prove much about the "ecology".

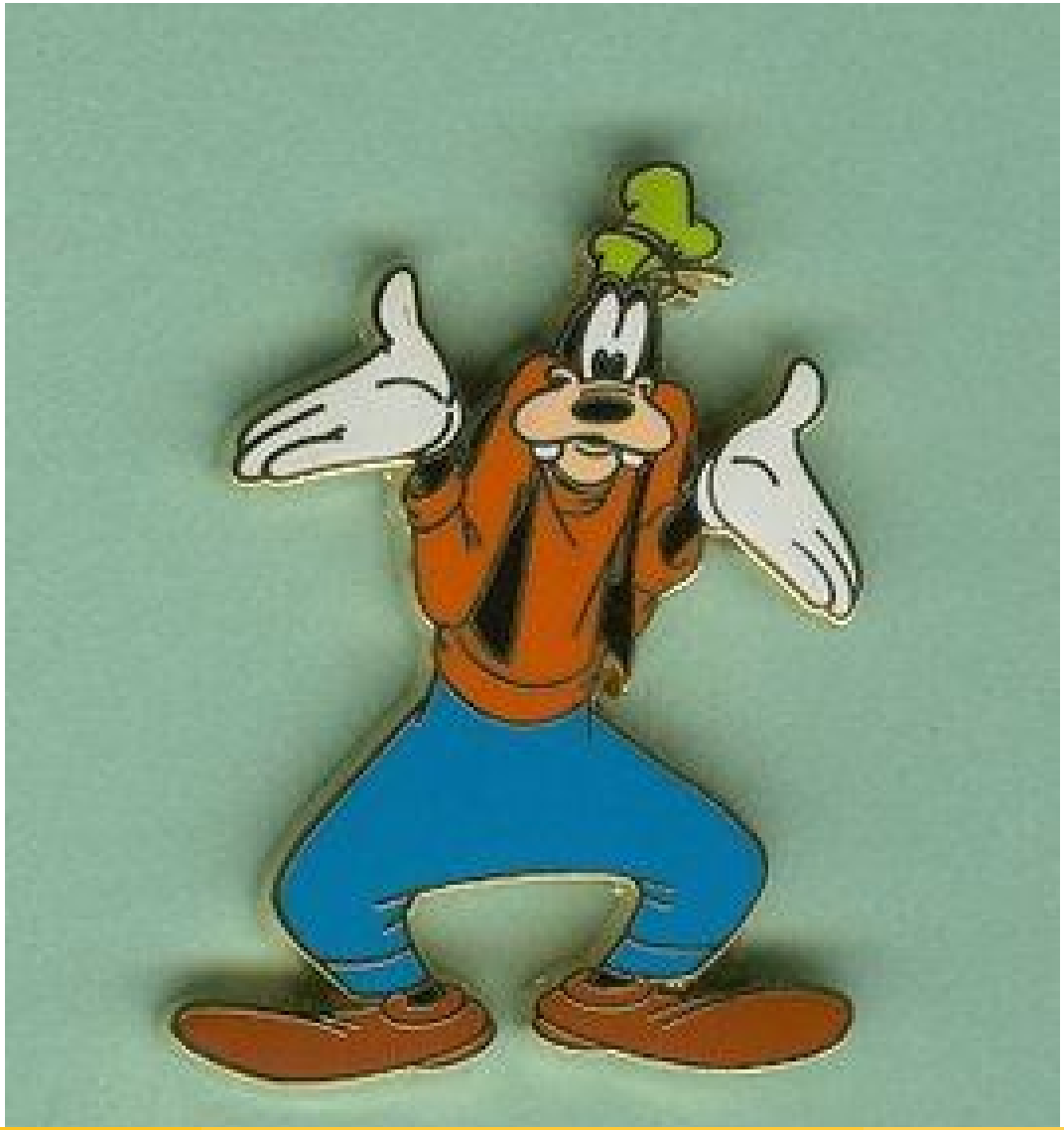
It's a first step.

- First had to show that corpus techniques even work for this social science problem.
- Only allows us to investigate related concepts.
  - What about the goal of investigating the community of interested parties?
- The technique does not use any structure in the texts.
  - But can the structure tell us something about the community?

And yeah, there are a quadrillion other uses for this stuff.

- Question-answering systems.
- Summarization.
- Stock market prediction (already done with Twitter).
- Political campaigning.
- Social psychology.
- ...

Big Data: solves everything and/or nothing



So next time, we'll talk about a way of making data more useful:

annotation