You go with the sequences you got.

Grammar-based approaches to opinion mining: Part 4 (ESSLLI 2013)

Asad Sayeed

Uni-Saarland

On the menu

Supervised and/or sequence-based techniques.

- Opinion source identification techniques.
- Opinion target challenges.

Q: So what was opinion mining again?

A: Let's go back to the beginning

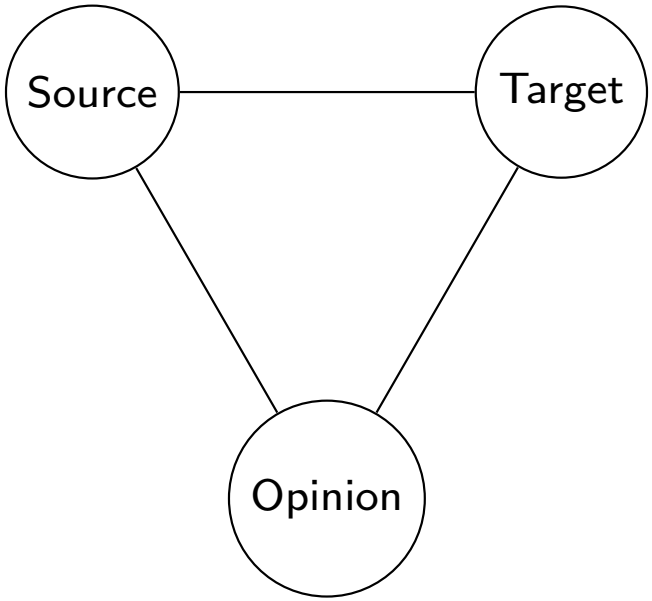
A non-definition

An *opinion source* holds an *opinion/sentiment* about an *opinion target*.

where

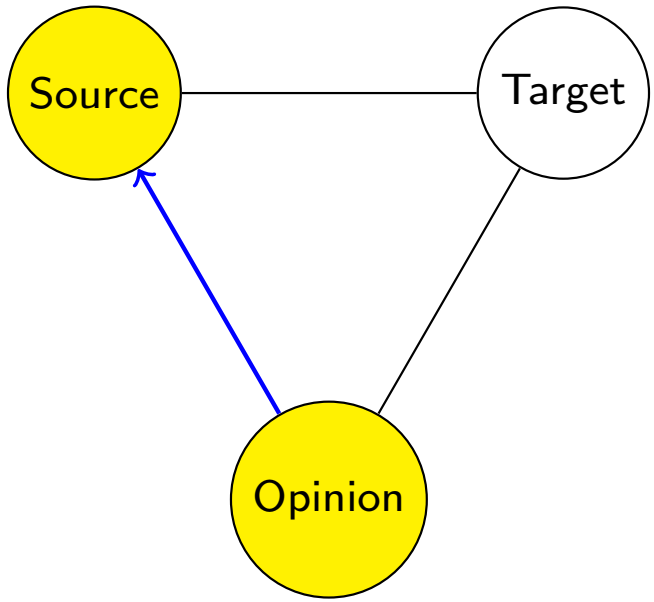
- The opinion source is the expressor/holder of the opinion (person/org).
- The opinion target is the entity or topic which the opinion is about.
- The opinion itself is generally represented as a polarity (or, more rarely, a more complex multidimensional construct).

They form a sort of triangle.



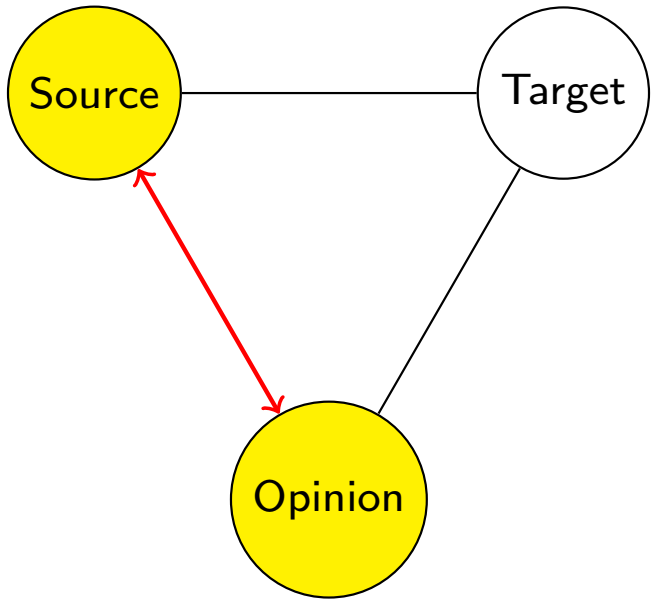
The presence of one is evidence for the other two.

They form a sort of triangle.



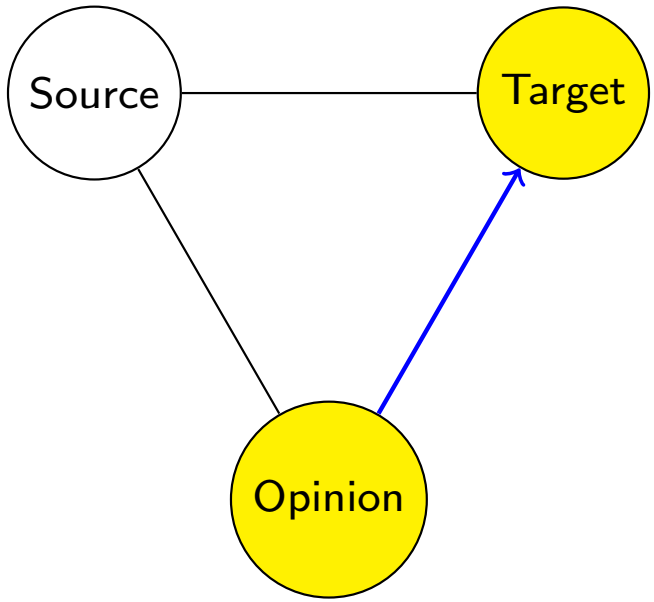
The opinion must exist for an opinion source to exist.

They form a sort of triangle.



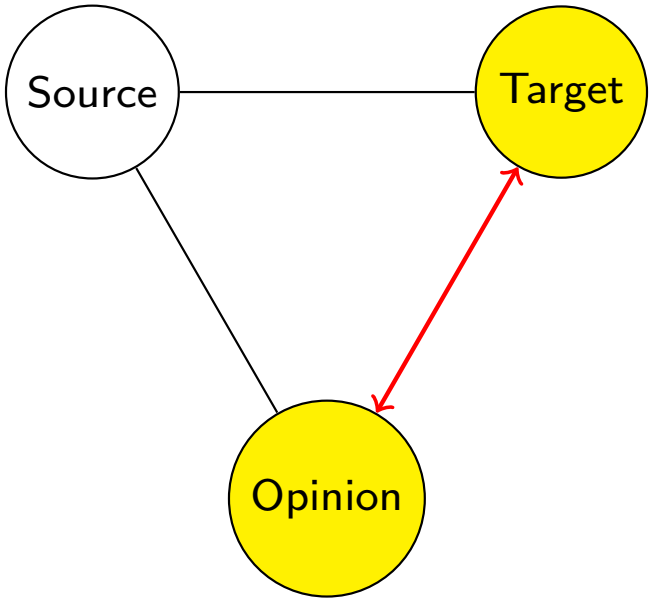
And if you can guess a source, you might find the opinion.

They form a sort of triangle.



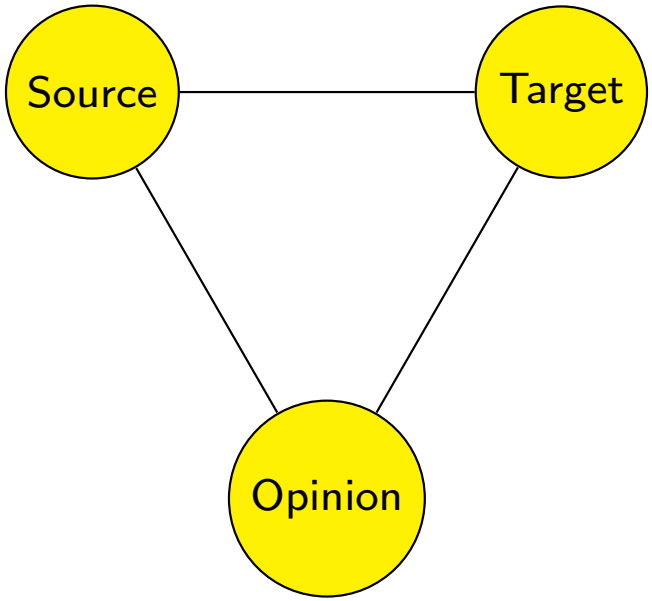
The opinion is generally about something.

They form a sort of triangle.



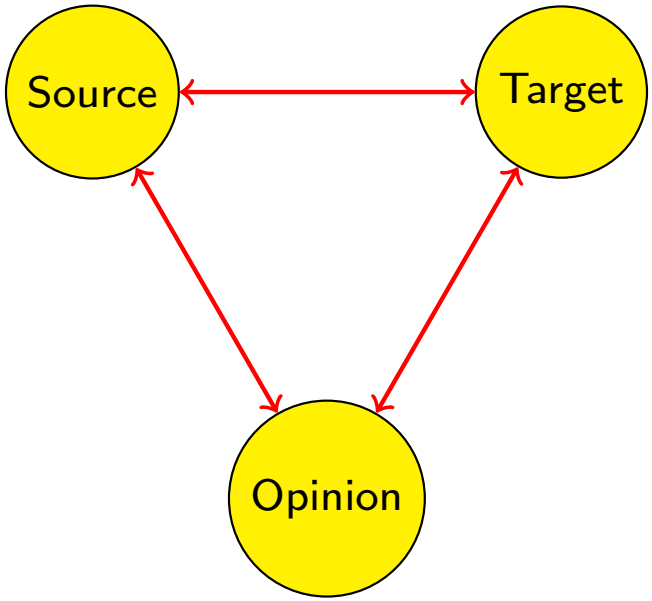
And a controversial topic might be evidence for an opinion.

They form a sort of triangle.



The “holy grail” is to identify them all together. . .

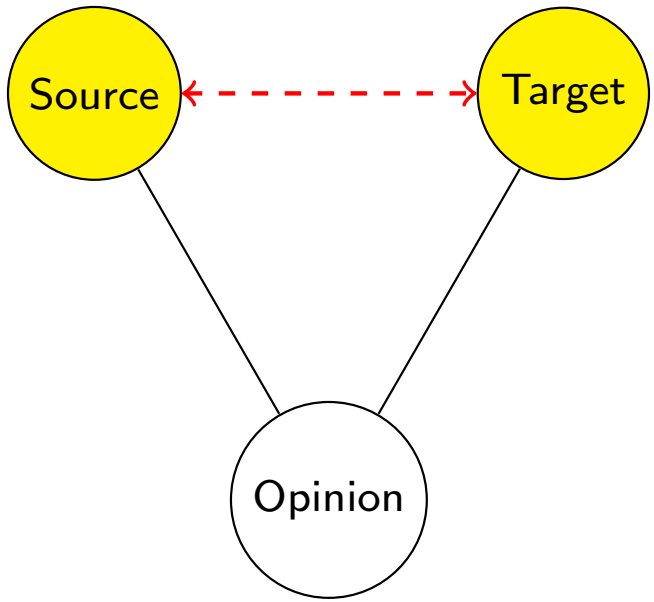
They form a sort of triangle.



. . . and to use any one as evidence for the other.

But there's one possibility that I left out.

The opinion mining triangle.



Sources and targets without (direct) help from opinion?

If we can eliminate the opinion. . .

. . . we could still track large-scale opinion trends.

- Reduce our need for "deep" NLP.
- Utilize information retrieval/extraction entity-detection methods $\rightarrow$ process larger data faster.

Essentially, it boils large-scale sentiment analysis down to a "relation-mining" problem.

Living without grammar? Sure!

Source-target relation mining in Sayeed et al. (2010):

- Yeah, IT business press again.
- Take into account pragmatic opinion ("opinionated acts"):

> And like other top suppliers to Wal-Mart Stores Inc., BP has been involved in a mandate to affix radio frequency identification tags with embedded electronic product codes to its crates and pallets. (ComputerWorld, January 2005).

- Goal: identify all source-target pairs at a paragraph level.

Living without grammar? Sure!

The data:

- 77,277 articles from ComputerWorld – person/org entities discovered by BBN IdentiFinder 3.3 (quite outdated, but suffices).
- List of IT innovations (as in the crowdsourcing stuff we talked about earlier).
- "Expert annotation" (two annotators—me and an undergrad collaborator) digging out source-target pairs by paragraph.

Then it just becomes a classification exercise.

You just need a lot of dimensions.

Everything boils down to features, with or without grammar.

- Unigram counts of words in candidate source-target pair's paragraph.
- Co-occurrence counts of how often source-target pair appear together within particular ranges of text (5, 25, 100, 500 words apart).
- Co-occurrence counts of source-target pair from Pitt subjectivity lexicon at 5, 25, 100, 500 words apart.

With some statistical feature reduction/filtering, we get a vector space of 8000 dimensions.

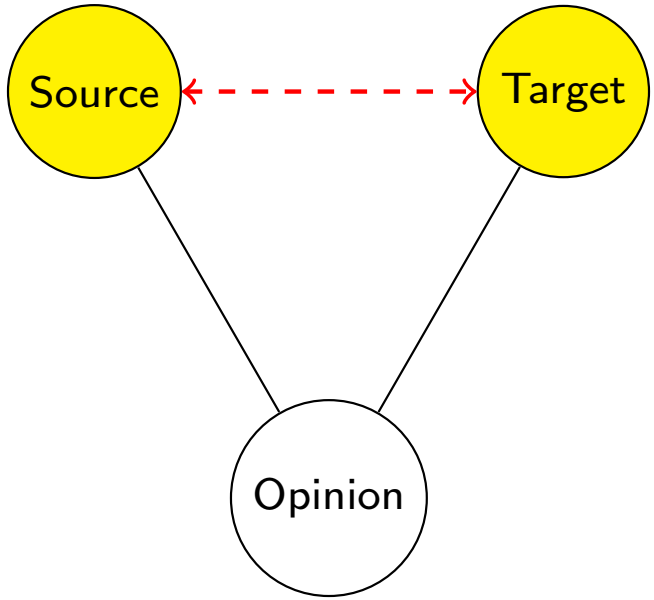
Then classify!

Support vector machine (SVM) classifier with something called a "radial basis function" kernel.

- Approx 1000 relations, slight majority of valid source-target relations.
- Precision and recall 0.72 and 0.76 (vs. random baseline of 0.60 and 0.53). Accuracy of 0.69 (vs 0.52).
- So better than random – possibly better enough to investigate large scale trends.
- Most importantly: best features in model came from the subjectivity lexicon, and strongly tended to be domain-related words.

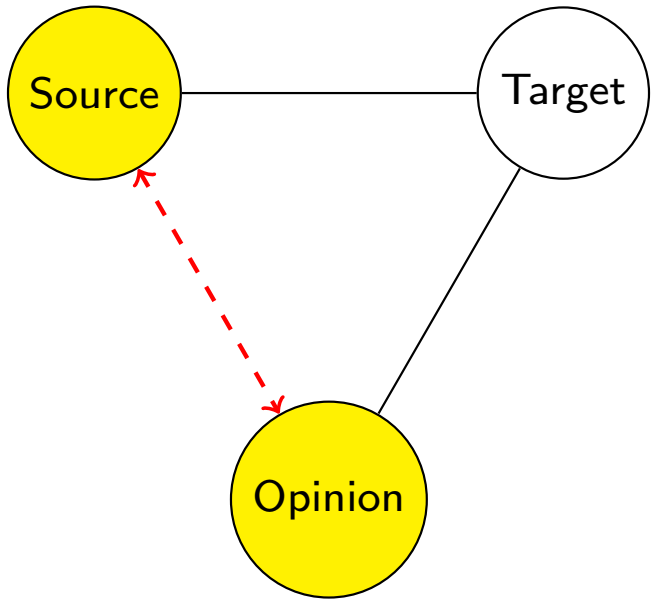
That last part tells us that we really are identifying opinions, but...

That's pretty much as far as we can go.



We can find out *that* opinions were expressed, but not how and why.

Motives depend on the source.



And so we need to know what it is that sources expressed.

Using CRFs and pattern extraction: Choi et al. (2005)

Another information extraction approach.

- Conditional random fields – model source identification as a problem of tagging word sequences based on features.
- Pattern extraction via AutoSlog – lexico-syntactic patterns based on predefined heuristics (Riloff et al., 1996) applied to surrounding contexts.
  - Statistical enhancement to AutoSlog to assign the probability of extracting a source to each pattern.

The overall method is supervised.

It works by labelling relevant sequences.

The output is not **directly** opinion sources.

- Tradition from HMM-based named-entity detection and other chunking schemes.
- Task: label multi-word sequences of opinion sources with start/middle/end markers—or irrelevant.
- Machine learning: classification problem, just like POS tagging, but with markers.

Sequence labelling techniques tend to work this way.

How do sources stand in relation to their corresponding opinions?

Choi et al. (2005) mention these cases:

Taiwan-born voters favoring independence. . .
According to **the report**, the human rights record in China is horrendous.
International officers believe that the EU will prevail.

First hand/direct sources of opinion.

International officers said **US officials** want the EU to prevail.

Indirect source of an opinion – "International officers"
Direct source – "US officials".

How do they decide on the features for the CRF?

Choi et al. considered these properties of opinion sources:

- Most sources of opinion are noun phrases.
- Sources should be semantic entities that can bear/express opinion.
- Should have direct relation to an opinion expression.
  - This is what makes this task different from named entity recognition!
  - Choi et al. claim features need to take into account sentence structure.

So what are these features anyway?

Quite a variety, including:

- Capitalization features – guess why you'd want this.
- Part-of-speech tags within a +/- 2-word window for NP segmentation.
- Opinion lexicon features within a +/- 1-word window.
  - From the Pitt opinion lexicon and from annotations in the training data.

These are the "flat" features...

. . . but they also use features based on grammatical structure.

Dependency tree features:

- Based on Collins (1999) dependency parser.
- Two steps, after parsing a sentence:
  1. Syntactic chunking – "flatten" tree by grouping constituents by grammatical role.
  2. Identify opinionated chunks as those that contain an opinion word.
- Then the features in the model consist of e.g. the grammatical role of the chunk to which a word belongs, parent's role, consituent boundaries, etc.

Semantic features also play an, ahem, role.

They include a few semantic features derived from Sundance shallow parser (Riloff, 2004).

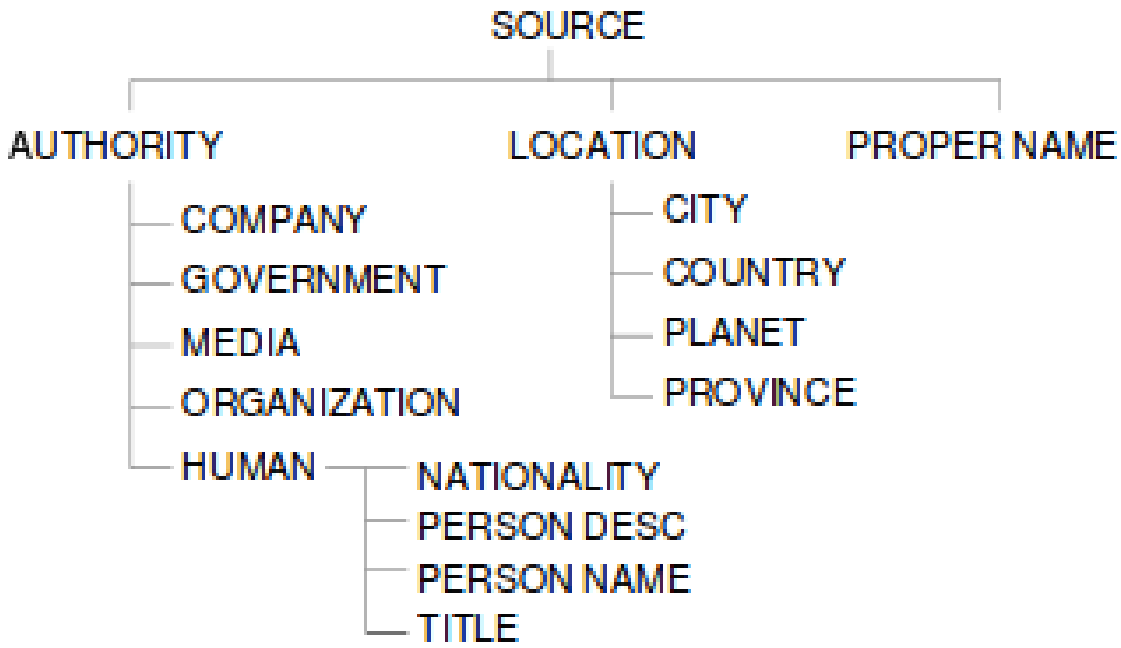


Figure 1: The semantic hierarchy for opinion sources

And this is where they integrate the AutoSlog patterns.

As features for the labelling CRF. (Remember that this is all per-word.)

- What a pattern looks like:
  - "<*subj*> *complained*"
  - The *subj* is the source whose identification is triggered by the presence of *complained*.
- How to CRFize this?
  - Source pattern features – does a word activate a particular extraction pattern? e.g. "complained" triggers "<*subj*> *complained*"
  - Source extraction features – is the word extracted by the pattern? e.g. "Chirac" in "Chirac complained".
  - Frequency and probability features thereof.

So does it work?

They developed with 135 MPQA documents and ran 10-fold cross validation on the other 400. Three evaluation metrics:

- Overlap match (OL) – "lenient" measure, how often do *any* words in the system output match the reference/gold standard?
- Head match (HM) – how often do heads of system-extracted entities match the reference data?
- Exact match (HM) – exactly what it says, strict.

Then precision, recall, F-measure based on these criteria.

There are baselines.

And they have three:

- Baseline 1: Label as source all phrases that can be classified as a member of plausible semantic categories (e.g. human, media, company).
- Baseline 2: Label as source all NPs that occur in compliance with a small list of contextual criteria (e.g. follows "according to").
- Baseline 3: Combine baselines 1 & 2 (conjunction/intersection).

Some additional conditions.

- Experiment with just extraction patterns alone.
  - Apply learned patterns to test data and accept whatever matches.
- Feature induction for the CRF (CRF-FI).
  - Technique to allow CRF training to automatically generate "conjunctive features".
  - Conjunctive features: features whose value is the conjunction of multiple features in the data.

And now, the results. . .

		Recall	Prec	F1
Baseline-1	OL	77.3	28.8	42.0
	HM	71.4	28.6	40.8
	EM	65.4	20.9	31.7
Baseline-2	OL	62.4	60.5	61.4
	HM	59.7	58.2	58.9
	EM	50.8	48.9	49.8
Baseline-3	OL	49.9	72.6	59.2
	HM	47.4	72.5	57.3
	EM	44.3	58.2	50.3
Extraction Patterns	OL	48.5	81.3	60.8
	HM	46.9	78.5	58.7
	EM	41.9	70.2	52.5
CRF: basic features	OL	56.1	81.0	66.3
	HM	55.1	79.2	65.0
	EM	50.0	72.4	59.2
CRF: basic + IE pattern features	OL	59.1	82.4	68.9
	HM	58.1	80.5	67.5
	EM	52.5	73.3	61.2
CRF-FI: basic features	OL	57.7	80.7	67.3
	HM	56.8	78.8	66.0
	EM	51.7	72.4	60.3
CRF-FI: basic + IE pattern features	OL	60.6	81.2	69.4
	HM	59.5	79.3	68.0
	EM	54.1	72.7	62.0

Table 1: Source identification performance table

The best results come when they use ALL the features.

But there's still a question of diminishing returns.

- They only get a couple of percentage points by adding the IE pattern features to the CRF-FI model.
- But who knows how critical that few might be?
- **Something to remind yourself**: hard to compare these results with anything that isn't this exact task on the MPQA.
  - True for everyone. Remember: task-dependence in resource construction!

What do the errors tell us?

For example, they think that this is a false positive:

> Perhaps this is why **Fidel Castro** has not spoken out against what might go on in Guantanamo.

They claim that their system incorrectly identifies Fidel Castro as a source.

- Does not recognize negation properly.
- But does that mean Fidel Castro hasn't uttered an opinion. . . ?

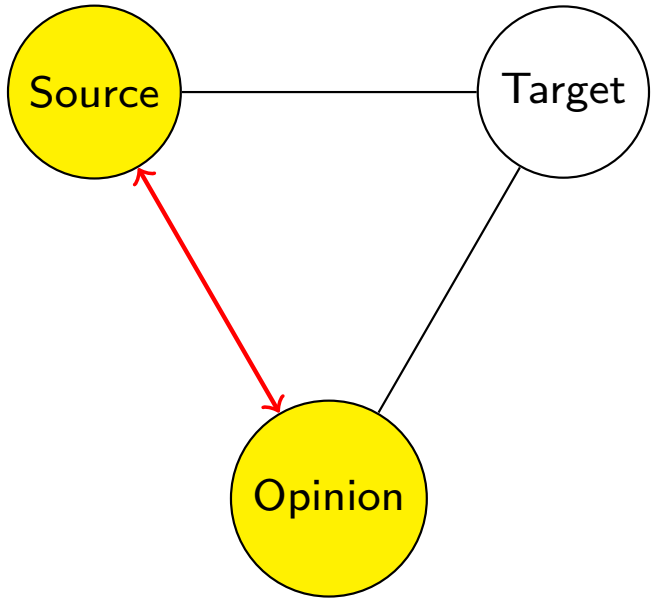
What do the errors tell us?

A clear false negative:

In particular, **Iran and Iraq** are at loggerheads with each other to this day.

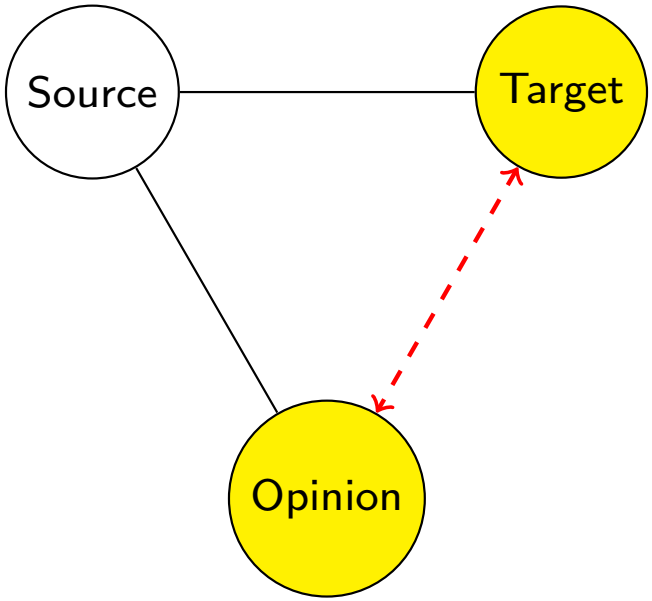
System didn't know the idiomatic expression "at loggerheads".

Thus, syntax and semantics can help identify source entities. . .



. . . by keeping track of patterns that also identify the opinion.

Can we do the same for targets?...



...by keeping track of patterns that also identify the opinion.

But targets are not quite the same as sources.

- Sources are often heralded by the opinion expression in close proximity.
  - "R2D2 believes that..."
  - "According to the Jedi Council..."
- But there's a greater variety of target contexts.
- Targets are often inanimate, abstract, etc.

Faraway are the targets.

From the MPQA:

In his view, Chien went on, the main purpose of Bush's current Asia visit is to promote world peace and security. "Bush hopes to take advantage of his trip to forge a consensus with U.S. allies in the Asia-Pacific area on his administration's fight against terrorism and the proliferation of nuclear weapons. . .

- MPQA annotates "consensus" as a sentiment-word about quite a long constituent!
- The constituent is separated from the sentiment word via a preposition.

CRFs to the rescue again

Jakob and Gurevych (2010): reviews and domain adaptation.

- Product review data sets.
- Attempt to apply models from one domain to another domain.
- CRFs with simple grammatical features from dependency paths.

What does the data look like?

Table 1: Dataset Statistics

	movies	web-services	cars	cameras
Documents	1829	234	336	179
Sentences	24555	6091	10969	5261
Tokens / sentence	20.3	17.5	20.3	20.4
Sentences with target(s)	21.4%	22.4%	51.1%	54.0%
Sentences with opinion(s)	21.4%	22.4%	53.5%	56.1%
Targets	7045	1875	8451	4369
Target types	865	574	3722	1893
Tokens / target	1.21	1.35	1.29	1.42
Avg. targets / opinion sent.	1.33	1.37	1.51	1.53

What features go into their model?

- Token features (tk) – just the word at current position.
- POS tag features (pos) from Stanford tagger.
- Short dependency path (dLn) – representing the presence of a *direct* dependency relation to an opinion word.
- Word distance (wrdDist) – label tokens in closest NP to an opinion expression with the distance to that expression, allows for multiword targets.
- Opinion sentence (sSn) – is word inside a sentence bearing an opinion expression.

Baseline: they have a nontrivial one

- Zhuang et al. (2006) on movie reviews: J&G's reimplementation.
  - This model extracts shortest paths from the dependency graph of a sentence – allows indirect links.
  - These paths connect opinion words to targets – use to detect paths in test data.

Table 2: Single-Domain Extraction with Zhuang Baseline

Dataset	Precision	Recall	F-Measure
movies	0.663	0.592	0.625
web-services	0.624	0.394	0.483
cars	0.259	0.426	0.322
cameras	0.423	0.431	0.426

So does Jakob and Gurevych (2010) work on a single domain?

CRF experiments run at 10-fold cross validation.

Table 3: Single-Domain Extraction with our CRF-based Approach

	movies			web-services			cars			cameras		
Features	**Prec**	**Rec**	**F-Me**	**Prec**	**Rec**	**F-Me**	**Prec**	**Rec**	**F-Me**	**Prec**	**Rec**	**F-Me**
tk, pos	0.639	0.133	0.220	0.500	0.051	0.093	0.438	0.110	0.175	0.300	0.085	0.127
tk, pos, wDs	0.542	0.181	0.271	0.451	0.272	0.339	0.570	0.354	0.436	0.549	0.375	0.446
tk, pos, dLn	0.777	0.481	0.595	0.634	0.380	0.475	0.603	0.372	0.460	0.569	0.376	0.453
tk, pos, sSn	0.673	0.637	0.653	0.604	0.397	0.476	0.453	0.180	0.257	0.398	0.172	0.238
tk, pos, dLn, wDs	**0.792**	0.481	0.598	0.620	0.354	0.450	0.603	0.389	0.473	0.596	0.425	0.496
tk, pos, sSn, wDs	0.662	0.656	0.659	0.664	0.461	0.544	0.564	0.370	0.446	0.544	0.381	0.447
tk, pos, sSn, dLn	0.791	0.477	0.594	0.654	0.501	0.568	0.598	0.384	0.467	0.586	0.391	0.468
tk, pos, sSn, dLn, wDs	0.749	**0.661**	**0.702**	**0.722**	**0.526**	**0.609**	**0.622**	**0.414**	**0.497**	0.614	**0.423**	**0.500**
pos, sSn, dLn, wDs	0.672	0.441	0.532	0.612	0.322	0.422	0.612	0.369	0.460	**0.674**	0.398	0.500

Take-home message #1: dependency path features help overcome problems of distance in a sequence-based model.

Take-home message #2: complex/long dependency paths are brittle.

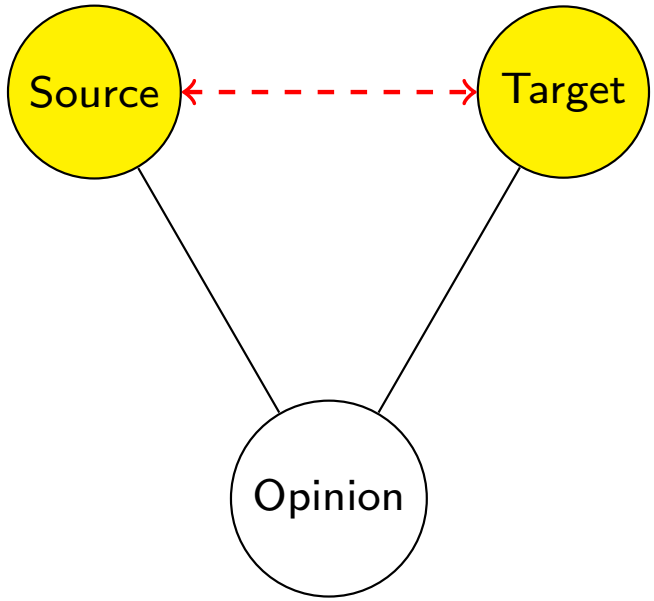
(The cross-domain part just for fun.)

Table 5: Cross-Domain Extraction with our CRF-based Approach

Training	Testing: web-services Pre	Rec	F-Me	movies Pre	Rec	F-Me
web-services	-	-	-	0.560	0.339	0.422
movies	**0.565**	0.219	0.316	-	-	-
cars	0.538	0.248	0.340	0.642	0.382	0.479
cameras	0.529	0.256	0.345	0.642	0.408	0.499
movies + cars	0.554	0.249	0.344	-	-	-
movies + cameras	0.530	**0.273**	**0.360**	-	-	-
movies + cars + cameras	0.562	0.250	0.346	-	-	-
cars + cameras	0.538	0.254	0.345	0.641	0.395	0.489
web-services + cars	-	-	-	**0.651**	0.396	0.492
web-services + cameras	-	-	-	0.642	**0.435**	**0.518**
web-services + cars + cameras	-	-	-	0.639	0.405	0.496
	cars Pre	Rec	F-Me	**cameras** Pre	Rec	F-Me
web-services	0.391	0.277	0.324	0.505	0.330	0.399
movies	0.512	0.307	0.384	0.550	0.303	0.391
cars	-	-	-	0.665	0.369	0.475
cameras	**0.589**	0.384	**0.465**	-	-	-
cameras + movies	0.567	**0.394**	**0.465**	-	-	-
cameras + web-services	0.572	0.381	0.457	-	-	-
movies + web-services	0.489	0.327	0.392	0.553	0.339	0.421
movies + cars	-	-	-	0.634	0.376	0.472
web-services + cars	-	-	-	**0.678**	0.376	**0.483**
web-services + movies + cars	-	-	-	0.635	**0.378**	0.474
movies + web-services + cameras	0.549	0.381	0.450	-	-	-

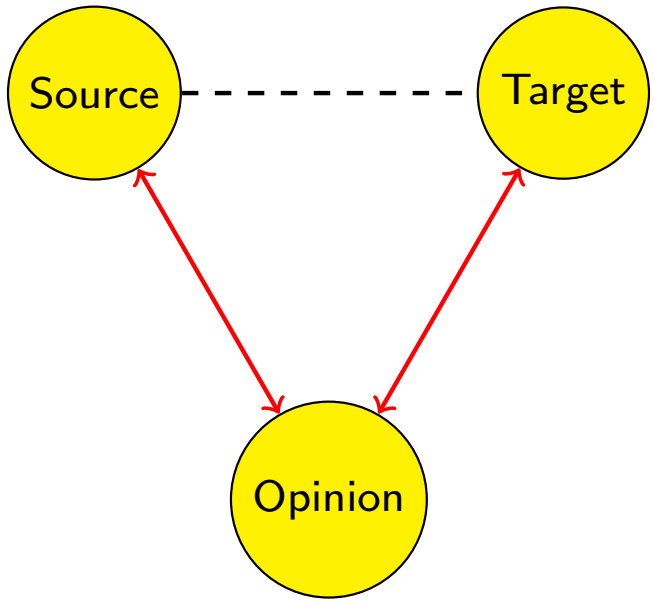
FYI: (Their use of) the Zhuang baseline is by comparison extremely low.

So, about those sources and targets.



We can sort of find these without involving the opinion.

So, about those sources and targets.



But what if we do involve the opinion?

Let's involve semantics.

Kim and Hovy (2006): use FrameNet.

1. Label opinions
   - Crafted dictionary of verbs and adjectives with positive/negative/neutral opinion.
   - Obtain related semantic frames from FrameNet.
2. Label semantic roles in corpus using existing role labelling algorithms.
3. Map semantic roles (Agent, Patient, etc) to opinion roles (Holder, Topic).
   - The role-mapping is by hand.

The usual question: does it work?

Sort of.

- Two humans annotated 100 newswire sentences.
- Baseline: use dictionary to identify opinion words.

Table 5. Opinion-bearing sentence identification on Testset 2. (P: precision, R: recall, F: F-score, A: Accuracy, H1: Human1, H2: Human2)

	P (%)	R (%)	F (%)	A (%)
H1	56.9	67.4	61.7	64.0
H2	43.1	57.9	49.4	55.0

Table 6: Results of Topic and Holder identification on Testset 2. (Sys: our system, BL: baseline)

		Topic			Holder		
		P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
Sys	H1	64.7	20.8	31.5	47.9	34.0	39.8
	H2	58.8	7.1	12.7	36.6	26.2	30.5
BL	H1	12.5	9.4	10.7	20.0	28.3	23.4
	H2	23.2	7.1	10.9	14.0	19.0	16.1

But there's one way in which we haven't exploited the grammar.

Actually learning to identify complex structures!

Which we'll talk about in the next part. . .