Recognizing Holders, Targets, and Topics

George Cooper and Scott Simpson

Definitions

- **Source/Subject/Opinion Holder:** The individual or entity that holds the opinion.
- **Opinion Expression:** a word that indicates that an opinion is being expressed
- **Target/Topic:** The the real-word object, event, or abstract entity to which an opinion refers in the context of a particular discussion.

John adores Marseille *and visits often*

MPQA Corpus

- Contains 535 documents
- Consists of news articles
- Manually annotated with opinion-related information
- Annotations include opinion sources

Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan

Problems/Goals addressed

- Goal: automatically identify the sources of opinions
- Critical for opinion-oriented question-answering systems and opinion-oriented summarization systems

Methodology

- Considered two different learning-based methods
- Semantic tagging via Conditional Random Fields
- Semantic tagging via Extraction Patterns

Semantic Tagging via Conditional Random Fields

Conditional Random Fields (CRF)

Use the IOB scheme to convert the task of "chunking" into a sequence tagging task.

[International officers] believe that the EU will prevail.

B I O O O O O O

Features

Capitalization Features

- Whether the word is all capital letters
- Whether the word begins with a capital letter

Part of Speech Features

- POS of the current token
- POS of the neighboring tokens in [-2, +2] window

Opinion Lexicon Features

- Whether the current token is in the opinion lexicon
- Whether the neighboring tokens in [-1, +1] window are in the opinion lexicon
- Opinion subclass (e.g. “moderately subjective”, “judgments”)

Dependency Tree Features

- Grammatical role (e.g. subject, object) of the current word's chunk
- Grammatical role of the previous word's chunk
- Whether the parent chunk of the current word's chunk includes an opinion word
- Whether the current word's chunk is in an argument position with respect to the parent chunk
- Whether the current word represents a constituent boundary

Semantic Class Features

- Individual words are labeled with semantic classes supplied by the Sundance shallow parser.
- **Classes:** authority, government, human, media, organization_or_company, proper_name, and other (classes that cannot be sources)

Induced Features

- Any helpful conjunctions of features are added (addresses CRF limitation)

Semantic Tagging via Extraction Patterns

AutoSlog

- **AutoSlog** - A supervised learning algorithm for pattern extraction generation.

"President Jacques Chirac frequently complained about France's economy."

Extraction Pattern: <subj> *complained*

Extracted Text: *"President" "Jacques" "Chirac"*

AutoSlog-SE

- An augmented version of AutoSlog
- Heuristics applied to every NP
- Augmented with selectional restrictions constraining NP's
- Patterns applied to training corpus and statistics are gathered about extractions that match.

Extraction Pattern Features

- Frequency of the highest-frequency pattern that the current word activates
- Probability of the highest-probability pattern that the current word activates
- Frequency of the highest-frequency pattern that extracts the current word
- Probability of the highest-probability pattern that extracts the current word

Baselines

- **Baseline-1**: All NP's with an appropriate semantic category are sources
- **Baseline-2**: All NP's that meet ANY of the following conditions are sources:
  - <NP-subj> <opinion VP>
  - "according to" <NP>
  - <opinion word> <NP>'s
  - <opinion word> "by" <NP>
- **Baseline-3**: All NP's that satisfy both Baseline-1 and Baseline-2

Results

	Recall	Precision	F1
CRF: basic features	50.0	72.4	49.2
CRF: basic features + IE pattern features	52.5	73.3	61.2

Results

	Recall	Precision	F1
CRF: basic features	50.0	72.4	59.2
CRF-FI: basic features	51.7	72.4	60.3

Results

	Recall	Precision	F1
Baseline-3	44.3	58.2	50.3
Extraction Patterns	41.9	70.2	52.5
CRF-FI: basic features	51.7	72.4	60.3
CRF-FI: basic + IE pattern features	54.1	72.7	62.0

Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields

Niklas Jakob and Iryna Gurevych

Problems/Goals addressed

- Extract opinion targets from user-generated online discourse.
- Existing annotated data from three domains:
  - Internet Movie Database (IMDb)
  - epinions.com
  - Blogs about digital cameras and cars
- Approach evaluated within each domain and cross-domain.

Example Annotated Sentence

While none of the **features**[1] are ***earth-shattering***[2], **eCircles**[1] does provide a ***great***[2] place to keep in touch.

[1] Underlined words denote opinion targets.

[2] Italicized words denote opinion expressions

Methodology

- Using the IOB scheme, the authors convert the task of "chunking" into a sequence tagging task.
- The authors model the problem as an IE task using CRF.

Features

- The text of the current token
- POS of the current token
- Whether a direct path exists in the dependency parse of the sentence from the current token to an opinion expression
- Whether the token is part of the closest noun phrase to an opinion expression
- Whether the token is part of an opinion sentence

Feature Comparison

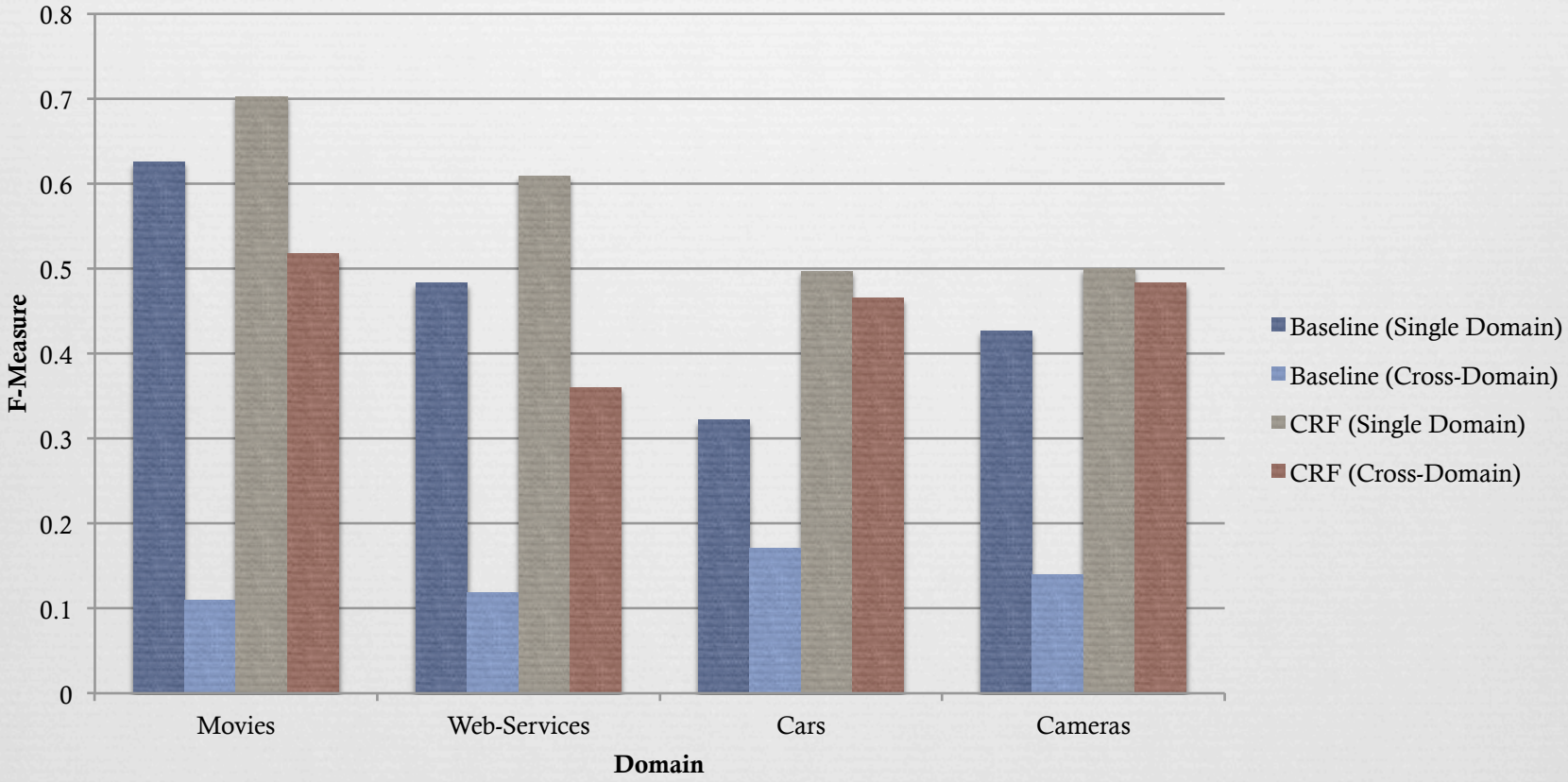
Identifying Sources (Cardie & Patwardhan)

- Capitalization
  - Whether the word is all capital letters
  - Whether the word begins with a capital letter
- Part-of-speech
  - *POS of the current token*
  - POS of the neighboring tokens in [-2, +2] window
- Opinion lexicon features
  - Whether the current token is in the opinion lexicon
  - Whether the neighboring tokens are in the opinion lexicon
  - Opinion subclass (e.g. "moderately subjective", "judgments")
- Dependency tree features
  - Grammatical role of the current word's chunk
  - Grammatical role of the previous word's chunk
  - *Whether the parent chunk includes an opinion word*
  - *Whether the current word is an argument of the parent chunk*
  - Whether the current word represents a constituent boundary
- Semantic class features
- Extraction pattern features
- Induced features

Extracting Targets (Jakob & Gurevych)

- The current token
- *POS of the current token*
- *Whether a direct path exists in the dependency parse of the sentence from the current token to an opinion expression*
- Whether the token is part of the closest noun phrase to an opinion expression
- *Whether the token is part of an opinion sentence*

Results


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
F-Measure
Movies
Web-Services
Cars
Cameras
Domain
Baseline (Single Domain)
Baseline (Cross-Domain)
CRF (Single Domain)
CRF (Cross-Domain)

Topic Identification for Fine-Grained Opinion Analysis

Veselin Stoyanov and Claire Cardie

Problems/Goals addressed

- Create an annotated corpus of topic information.
- Build an automated system for identifying references to a common topic in a document.

Challenges

- Multiple potential topics for each opinion expression:

  Example: *Al thinks that **the government** should **tax gas** more in order to curb **CO_2 emissions**.*

*bold denotes potential topics

- Opinion topics not always explicitly mentioned:

  Example: *John identified the violation of Palestinian human rights as one of the main factors.*

  [Topic: Israeli-Palestinian Conflict]

Methodology

- Treat problem as a topic coreference resolution task.
- Each pair of topics is separately classified as being co-referent or not.
- Extend MPQA corpus with manual annotations that encode topic information.
- Train and test a classifier using extended corpus.

Inter-Annotator Agreement

	alpha
All opinions	0.5476
Sentiment opinions	0.7285
Strong opinions	0.7669

Features

Features: Positional

- Opinions in same sentence?
- Opinions in same paragraph?
- Opinions in consecutive sentences?
- Opinions in consecutive paragraphs?
- Number of sentences separating opinions
- Number of paragraphs separating opinions

Features: Lexico-Semantic

- The cosine similarity of the tf-idf weighted vectors of the terms contained in the two spans
- Whether the two spans have any words in common
- Whether the two spans contain coreferent NPs (according to "simple rule-based coreference system")
- Whether the two spans contain entities that can be considered aliases of each other

Features: Opinion

- Whether both opinions have the same holder
- Whether both opinions have the same polarity
- Whether both opinions have the same holder but opposite polarities

Baselines

- **One topic:** All opinions are in the same cluster
- **One opinion per cluster:** Each opinion is its own cluster
- **Same paragraph:** One cluster per paragraph
- **Choi 2000:** One cluster per topic, as identified by the topic segmentation algorithm presented in Choi (2000)

Topic Span Identification

- **Sentence:** Topic span is whole sentence containing opinion.
- **Automatic:** Rule-based method for identifying the topic span. Rules dependent on syntactic constituent type of opinion expression, relying on parsing and labeling.
- **Manual:** Topic span marked by human annotator.
- **Modified Manual**: Returns the manually identified topic span only when it is within the sentence of the opinion expression. Returns opinion sentence when outside sentence boundary.

Results

	alpha
One topic	-0.1017
One opinion per cluster	0.2238
Same paragraph	0.3123
Choi	0.3734
Sentence	0.4032
Rule-based	0.4056
Modified manual	0.5134
Manual	0.6585

Conclusion

- All three papers address the question "Who thought what about what?"
- Useful for opinion-based QA systems and opinion summarization systems
- All three papers use machine learning algorithms and use opinion expressions.
- MPQA corpus common to two of the systems
- All three tasks have a lot of potential for new research