Can summaries help natural language processing?

Studying the effect of text summarization in opinion classification and other tasks

Horacio Saggion
TALN
Universitat Pompeu Fabra
Barcelona, Spain

Initial Words

- Thanks to Frederik Cailliau (Sinequa), Javier Couto (Syllabs), and to the project RPM2
- Motivation
  - Automatic Summarization is imperfect
  - Automatic Summarization by extraction continues to be the paradigm
  - Can we take advantage of automatic summaries (given their current standards) to accomplish natural language processing tasks?
- This talk presents a first look into the problem

Outline of the Talk

- Short overview of text summarization
- Short overview of summarization evaluation
- Summarization in natural language processing
  - Summarization-based Cross-document coreference
  - Summarization-based Opinion Classification
  - Discussion
- Final words

Text Summarization Overview

- A summary is a brief but accurate text representation of the content of a document or set of documents
- A summarization system should extract what is important in the source taking into account some input requirements and produce a textual summary matching some output requirements:
  - there are various summarization tasks!
- Most summarization approaches follow a "select and concatenate" approach – production of extracts
- Studies on text abstracting also exist but are for the time being limited
- Available summarization tools are sentence extraction systems

Text Summarization Overview

- Summarization by sentence extraction relies on a set of features to measure sentence relevance
- Features can be used to produce a score for each sentence and scores used decide which sentences are more important
- Features can also be used in a machine learning approach to create a classifier

Text Summarization Overview

- Statistical approaches
  - tf*idf statistics (Luhn, 1958)
  - cue-words (Edmudson, 1969; Paice, 1980)
  - title information (Edmundson, 1969)
  - centroid (Radev el al. 2000, Saggion & Gaizauskas, 2004)
  - position (Lin & Hovy, 1997)
  - graph-based methods (word, sentence connectivity) (Salton et al, 1999; Erkan & Radev, 2004; Mihalcea, 2004)
- Knowledge rich approaches
  - lexical resources (WordNet) and lexical cohesion (Barzilay & Elhadad, 1997; Benbrahim & Ahmad, 1994)
  - discourse organization theories (Marcu, 1998; Teufel & Moens, 2002)
  - information extraction & generation approaches (Oakes & Paice, 2001; Saggion & Lapalme, 2002)

Text Summarization Evaluation

- Intrinsic paradigm: check the content and quality of the summary in itself
  - Compare automatic summary with one (or more) ideal human written summaries
  - Based on comparison, a score is assigned to an automatic summary
  - Scores are aggregated per system and systems ranked
  - Metrics used: Coverage, ROUGE (Lin, 2004), Pyramid (Nenkova & Passoneau, 2004), Cosine (Donaway et al, 2000)
  - Grammaticality and responsiveness are also important
  - Intrinsic evaluation has been extensively studied in the DUC 2000-2007 and TAC 2008 – 2010 conferences (Over et al., 2007)

Text Summarization Evaluation

- Evaluation without models
  - Based on work by Louis & Nenkova (2009) we developed FRESA (Université d'Avignon & UPF) (Saggion et al., 2010; Torres-Moreno et al., 2010)
    - Summaries are compared to the source document(s) using a "divergence" and a score produced
    - Summarizers ranked based on the score
    - For some summarization tasks FRESA metrics produce rankings which correlate with human rankings
      - There are tasks where correlation can not be established

Text Summarization Evaluation

- Extrinsic paradigm: check if the summary can be used instead of the original (not abridged document) in a task
- Extrinsic evaluation is one way to check if your summaries are really helpful
- Some scenarios where it has been used:
  - SUMMAC evaluation framework: classification & QA (human)
  - Essay evaluation (human)
  - Assigning keywords to documents (human)
  - Information retrieval (machine)

Summarization-based Cross-document coreference

- Search for information about people or other entities (e.g. companies, locations) in huge text collections or on-line repositories on the Web
- Problem: names are highly ambiguous
- Cross-document coreference resolution is the task of identifying if two mentions of the same (or similar) name in different sources refer to the same individual
- Solving this problem is important in practical information access applications such as summarization and question answering

Summarization-based Cross-document coreference

- Web People Search Evaluation (Artiles el at, 2007; 2010)
  - Given a set of documents retrieved from a document collection in response to a person name query, a system has to organize the documents in clusters each referring to an individual
  - Text clustering can be used to solve this problem
    - Represent documents as vectors of terms and compare them
    - Put "similar" documents in the same cluster until a stop criteria is met
  - We use agglomerative clustering algorithm and test it using different conditions

Summarization-based Cross-document coreference

- Data for experimentation (1st Web People Search evaluation)
  - Training: 10 data points (person names) and 100 documents from the Web for each name
  - Testing: 30 data points (person names) and 100 documents from the Web for each name
  - Evaluation metrics: F-score (combination of clustering evaluation metrics Purity and Inverse Purity)
  - The dataset will contain things such as
    - "Donna Harman" – scientist, state agent, …
    - "George Foster" – actor, sportsman, …

Summary-based Cross-document Coreference

- Use agglomerative clustering with a similarity metric to measure "how close" two vectors are
  - Given a set of documents and a threshold
    1. put each document in a different cluster
    2. compare all clusters using a similarity metric
    3. the two most similar clusters are merged if their similarity is greater than a threshold (otherwise stop and return clusters)
    4. continue with step 2

Summarization-based Cross-document coreference

- Vector representations of text: word based, semantic-based (i.e., basic named entities: org, loc, per, date, address)
- Text analysis with off-the-shelf components from GATE (Cunningham et al, 2000)
- Similarity between texts: cosine
- Texts: full document or summary
- Summary: set of sentences "containing" the target named entity (not compression parameter here)

Summarization-based Cross-document coreference

- Results
  - Vectors of words extracted from documents or summaries performed equally well and vectors of named entities from documents or summaries performed worst
  - Further analysis of the use of semantic information was carried out
  - Using specific types of named entities provide improved results for both full documents and summaries (4 to 6 points improvement)
  - Using "Organization" with full document condition
  - Using "Organization" or "Person" with summary condition
- Previous studies have indicated that personal summaries worked better than full documents
- We decided to study the effect of different types of summaries in the task

Summarization-based Cross-document coreference

- Studying different types of summaries (and compressions) in combination with each type of information
- Text analysis: the GATE system
  - extraction of named entities of type: ***Org, Loc, Addr, Per, Date***
- Summarization tool: the SUMMA system (Saggion, 2008)
- Summarization features investigated:
  - position; cue feature; query feature; semantic feature
- Summaries created
  - lead summary; cue-based summary; semantic-based summary; query-based summary; semantic + query-based summary; cue + query-based summary
- Compressions: 10, 20, 30, 40, 50% of the text

Summarization-based Cross-document coreference

- We run around 900 experiments (combinations of summaries, compressions, semantic information) to see how the clustering performed
- Sem = semantic-based summary
- Qbased = query-based summary
- Lead = lead-based summary
- Cue = cue-based summary

Configuration	F-Score
Sem+Org+40	0.78
Sem+Org+50	0.78
Sem+Qbased+Org+50	0.78
Sem+Org+30	0.78
Lead+Org+50	0.77

Configuration	F-Score
Sem+Addr+10	0.61
Lead+Addr+20	0.61
Cue+Addr+20	0.60
Cue+Qbased+Addr+20	0.59
Cue+Qbased+Addr+10	0.58

Summarization-based Cross-document coreference

- Findings:
  - Summaries which are dense in named entities and contain the target entity perform on average as good as the best configuration using the full document
  - Lead-based summaries at 50% are close to best solution and not different from best
  - Organization type of information provides better overall performance
- Are summaries helping in this task?
  - Overall some summarization strategies performed as well as using the whole document
  - Although computation of summaries is not too expensive, they did not provide overwhelming advantage in clustering
  - The summarization space is quite big and it is possible that other summary types provide better results

Summarization-based Opinion Classification

- Work in collaboration with Elena Lloret (Universidad de Alicante) and Manuel Palomar (Universidad de Alicante)
- Opinion classification (Pang & Lee, 2005)
  - the rating-inference problem (what is the degree of the expressed opinion)
- We investigate what type of summary could be used for inferring the positivity or negativity of a given opinion

Summary-based Opinion Classification

Movie Reviews

Soul Survivors

Soul Survivors tells the story of four college-bound friends… Driving back from a creepy, gothic party, the quartet gets into a horrific car crash….

There's no captivating dialogue, no character chemistry exists anywhere… Soul Survivors is so awful I feel compelled to knock on doors and warn people about it.

Where is my friends house?

A young Northern Iranian schoolboy, Ahmed, wishes to return a school notebook to his classmate, who is threatened with expulsion from school if he does not bring in his notebook to school for tomorrow's class.

…but this is still great cinema, something American filmmakers would be wise to observe…. What more can you ask for?

Summary-based Opinion Classification

- Product/Company reviews
  - I ordered some suitcases on the 20th from www.thesportshq.com ….i got them the very next morning!!! the cases were great value for money, arrived super quick and I am very pleased with the quality. and service i recieved. would def shop again…
  - Let me give you a little bit of history, I have been with them for 23 years (with a few minor niggles, but nothing much to be concerned over), but the last 4 years the bank has hit the bottom in terms of customer service (the bane of my life) …My advice to you is steer well clear of the Abbey, they are worse than useless, they are unhelpful, arrogant and dangerous …

Summary-based Opinion Classification

- One key research questions is: what words (or features) can help an algorithm identify an opinionated text, the strength of the opinion, etc.?
  - Words which can more or less indicate an opinion: captivating, awful, great, recommend, very pleased, poor, unhelpful, arrogant, ...
  - Expressions which are rather unique: def shop again, what more can we ask for?, steer the wheel clear of... , nobody returned our calls,...

- Some works look at deriving or using words considered "opinionated"
  - Syntactic contexts ("interesting and useful" vs. "beautiful but boring") can give clues of positive and negative orientation (Hatzivassiloglou & McKeown, 1997)
  - The semantic orientation of words can be established by computing their pair-wise mutual information (PMI) to referents "excellent" and "poor" (Turney, 2002)

Summary-based Opinion Classification

- Word features: word root, part-of-speech
- Use SentiWordNet a lexical resource that associates to each word 3 numerical scores (obj, pos, neg) indicating how positive, negative, or objective is the given "word" (Esuli & Sebastiani, 2006)
- Sentiment features extracted from SentiWordNet: we compute the positivity, negativity, or neutrality of a word

Cat	WNT #	pos	neg	synonyms
a	1006645	0.25	0.375	good well
a	1023448	0.375	0.5	good unspoilt unspoiled
a	1073446	0.625	0.0	good
a	1024262	0.0	1.0	bad spoilt spoiled

Machine Learning Tool

- We use a machine learning library of Support Vector Machines algorithms. SVMs are particularly good for text classification tasks where the number of features is huge (Joakims, 1998)
- Instances are represented in a vector space (of features) and the algorithms tries to identify an hyperplane that separate positive from negative instances
- We use an implementation of SVMs provided by the GATE system
- The experiments consist on classifying the review using either the full review or a summary of the review

Summary-based Opinion Classification

- What type of summary and compression rate is more appropriate to predict the rating of a review
- Data for experimentation
  - One dataset of 89 bank reviews from the web – ratings between 1 and 5
  - The movie review dataset – different ratings available, we have used 3 class rating
    - 4 subsets with 1027, 902, 1307, 1770 reviews each
  - Evaluation metric:
    - Mean Square Error to capture deviation from the true class

Summary-based Opinion Classification

- Text summarization strategies: (Lloret & al, 2010)
  - Summarization features
    - Term frequency
    - Code Quality Principle (= noun phrase relevance)
    - Query relevance
    - Position (initial/final)
  - Text entailment technique for redundancy removal (Ferrandez et al, 2007)
  - Off-the-shelf sentence sentiment scoring (Balahur-Dobresco et al, 2009)
  - Compressions: 10, 20, 30, 40 & 50 %

Summary-based Opinion Classification

- Bank review dataset
  - Using word roots (bag of words) to represent each document/summary
  - Using combination of features to represent each document/summary
- Results
  - 160 conditions tested (summaries x compressions x type of representation); over 7,000 summaries
  - Many summary conditions performed worst than the full document condition
  - Some summary conditions performed close to full document performance

Summary-based Opinion Classification

- Summary classification results (MSE) using words

SUMMARY	10%	20%	30%	40%	50%
lead	3.10	3.00	3.10	3.30	3.10
final	2.74	3.00	**2.13**	2.64	**2.48**
qf	**2.49**	2.70	3.58	3.78	3.88
sent	3.89	3.16	3.03	2.90	2.66
generic-tf	3.21	**2.33**	**2.39**	**2.37**	**2.44**
generic-te+tf	3.39	3.23	**2.52**	**2.38**	**2.29**
generic-cqp+tf	3.01	3.34	2.61	3.17	3.03
generic-te+cqp+tf	2.70	2.93	3.00	3.10	2.71
qf-tf	**2.11**	**2.19**	**2.18**	**2.46**	**2.37**
qf-te+tf	**2.44**	**2.08**	**2.30**	**2.27**	**2.42**
qf-cqp+tf	2.83	3.00	2.70	**1.80**	**2.00**
qf-te+cqp+tf	**2.11**	**2.51**	**2.28**	**2.40**	**2.10**
sent-tf	2.83	**2.16**	**2.47**	**2.43**	**2.29**
sent-te+tf	3.20	2.80	**2.40**	2.69	2.71
sent-cqp+tf	3.01	3.27	2.62	3.21	3.10
sent-te+cqp+tf	2.69	3.21	3.46	2.90	2.93

Summary-based Opinion Classification

- Summary classification using combination of features

SUMMARY	10%	20%	30%	40%	50%
lead	2.63	2.62	2.72	2.86	2.60
final	3.20	2.66	**2.16**	**2.13**	**2.50**
qf	**2.45**	**2.53**	2.90	2.91	2.81
sent	3.69	2.87	**2.44**	**2.34**	**2.27**
generic-tf	3.21	2.68	**2.19**	**2.29**	**2.33**
generic-te+tf	3.98	2.90	2.79	3.06	2.64
generic-cqp+tf	3.37	3.45	3.31	3.19	2.84
generic-te+cqp+tf	3.02	2.87	2.83	3.18	**2.53**
qf-tf	**2.44**	**2.56**	**2.37**	**2.77**	**2.50**
qf-te+tf	3.21	**2.52**	**2.26**	2.62	2.72
qf-cqp+tf	2.73	3.48	**2.51**	**2.48**	**2.27**
qf-te+cqp+tf	2.60	2.83	**2.52**	**2.44**	**2.53**
sent-tf	3.02	**2.47**	**2.16**	**2.12**	**2.31**
sent-te+tf	2.99	3.07	**2.21**	2.93	2.64
sent-cqp+tf	3.54	2.81	**2.51**	**2.43**	2.70
sent-te+cqp+tf	2.77	**2.31**	**2.14**	2.82	**2.47**

Summary-based Opinion Classification

- Movie review dataset
  - Sentiment scale data set: 3 class & 4 class
    - we used the 3 class dataset
  - Two classification conditions: word root or SentiWordNet feature
- Preliminary results
  - Conditions tested = more than 20 different types of summaries x compression rates
  - Many summary conditions performed worst than the full document condition
  - Full documents work better when the classifier is trained with word-based information
  - Summaries performed differently in each subset

Summary-based Opinion Classification

- Word-based classification
  - Only in two subsets baseline summarizers performed as good as the full document
    - "Final" baseline (40-50% compression) in one subset
    - "Lead" baseline (50% compression) in another subset
- Sentiment-based classification
  - In four subsets there are summarization strategies performing better than full document
    - "Lead" (40-50% compression) and "Term Frequency" (with redundancy removal) (40% compression)

Summary-based Opinion Classification

- Are summaries helping in this task?
  - In the small set of bank reviews a number of summaries achieve a lower MSE compared to the full review, but we could not find evidence for statistical significance
  - In the bigger data set of movie reviews we found some baselines achieving lower MSE compared to the full review, but no across reviewers

Discussion

- Evaluation of text summarization is a controversial issue
- Current intrinsic evaluation of summaries relies on content-based metrics
- Extrinsic evaluation of summaries usually requires humans to perform specific tasks with the summary

Discussion

- We are studying a framework for evaluation where summarization system are tested based on a task carried out by a machine
  - We have carried out experimentation trying to answer the following questions
    - "Can a summary help predict the rating of a review?"
    - "Can a summary help identify the identity of a person?"
  - We have seen some summarization strategies performing as well as the full document
  - There are still factors to analyse such as amount of training data
  - We think there are possibilities for testing such a framework in other tasks such as question answering, information extraction, knowledge extraction

Can summaries help natural language processing?

May be…

Merci!

Questions?