The DLSIUAES Team's Participation in the TAC 2008 Tracks – Opinion Pilot

Alexandra Balahur, Elena Lloret,
Andrés Montoyo, Manuel Palomar

Universitat d'Alacant
Universidad de Alicante

Department of
Software and
Computing Systems

gPLSI

Research Group of Language
Processing and Information
Systems

Overview

- Task definition
- Objectives of participation
- Question processing
- Answer retrieval
- Summary generation
- Evaluation & discussion
- Conclusions & future work

Opinion pilot task definition

- Input - (opinion) questions from the TAC QA Track and the text snippets output by QA systems.
- Goal - produce short coherent summaries of the answers to the questions
  - from the text snippets themselves, or from the associated documents.
- Evaluation - readability and content (Nugget Pyramid Method)

Description of test data

- 25 topics
  - 22 with two questions
    - Usually asking positive/negative aspects on the topic
    - Comparisons among 2 objects
  - 3 with just one question
    - Only the positive or negative aspects of an entity
- Answer snippets – variable number
  - Correspondence between answer snippets and question not provided

Objectives of participation

- What is needed to build an MPQA system
- Difference to classical QA systems in question analysis & answer retrieval
- Test a general opinion mining system
- Test the relevance of different resources and techniques to these tasks
- Test importance of opinion strength to summarization

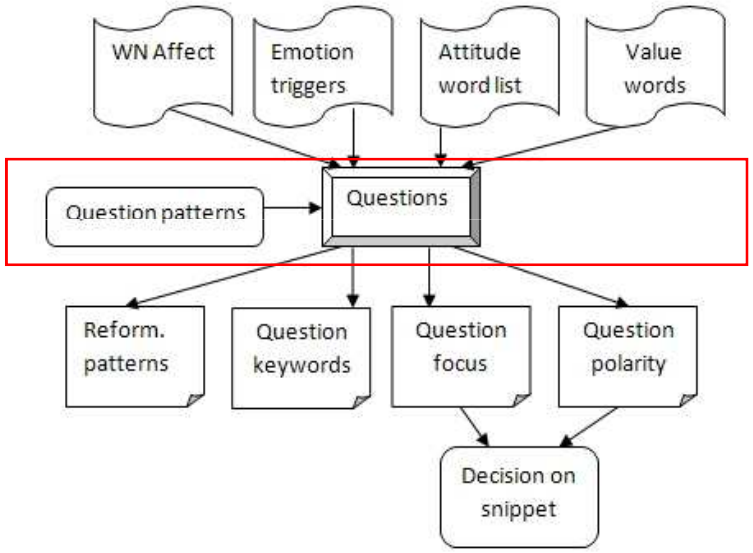
Question processing stage

- **Question patterns**
  - interrogation formula
  - opinion words.

Examples of rules for the interrogation formula

"What reasons" are:

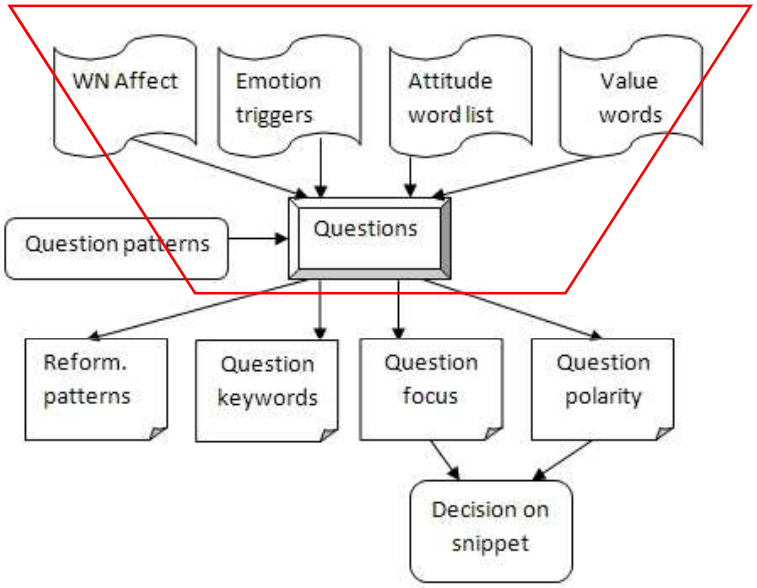
- *What reason(s) (.*?) for (not) (affect_verb + ing) (.*?)?*
- *What reason(s) (.*?) for (lack of) (affect_noun) (.*?)?*
- *What reason(s) (.*?) for (affect_adjective|positive|negative) opinions (.*?)?*



Question processing stage

Question polarity

- WordNet Affect (Strapparava and Valitutti, 2006) emotion lists
- the emotion triggers resource (fight, destroy, burn etc.) (Balahur and Montoyo, 2008)
- list of attitudes for the categories of criticism, support, admiration and rejection (em. triggers)
- two categories of value words (good and bad) - opinion mining system.



Words that denote human needs and motivations, whose presence triggers emotion.

Question processing stage

- **Question keywords**
  - filtering out stop words.
- **Question focus**
  - determining the gist of the question.
- **Output of the question processing stage:**
  - *reformulation patterns* (coherence to summaries) ,
  - *question focus*, *keywords* and the *question polarity* (->define several ***rules*** to make a correspondence between the question and the answer snippets on the further processing stage).

Correspondence rules

1. One question on the topic ⇒ retrieved snippet has same polarity as the question.
2. Two questions on the topic with different polarity ⇒ the snippets retrieved are classified according to their polarity.
3. Two questions with different focus and polarity ⇒ the snippets retrieved are classified according to their focus and polarity.
4. Two questions with the same focus and polarity ⇒ the order of the entities in focus both in the question and in the answer snippets is taken into account, together with a polarity matching between the question and the snippet.

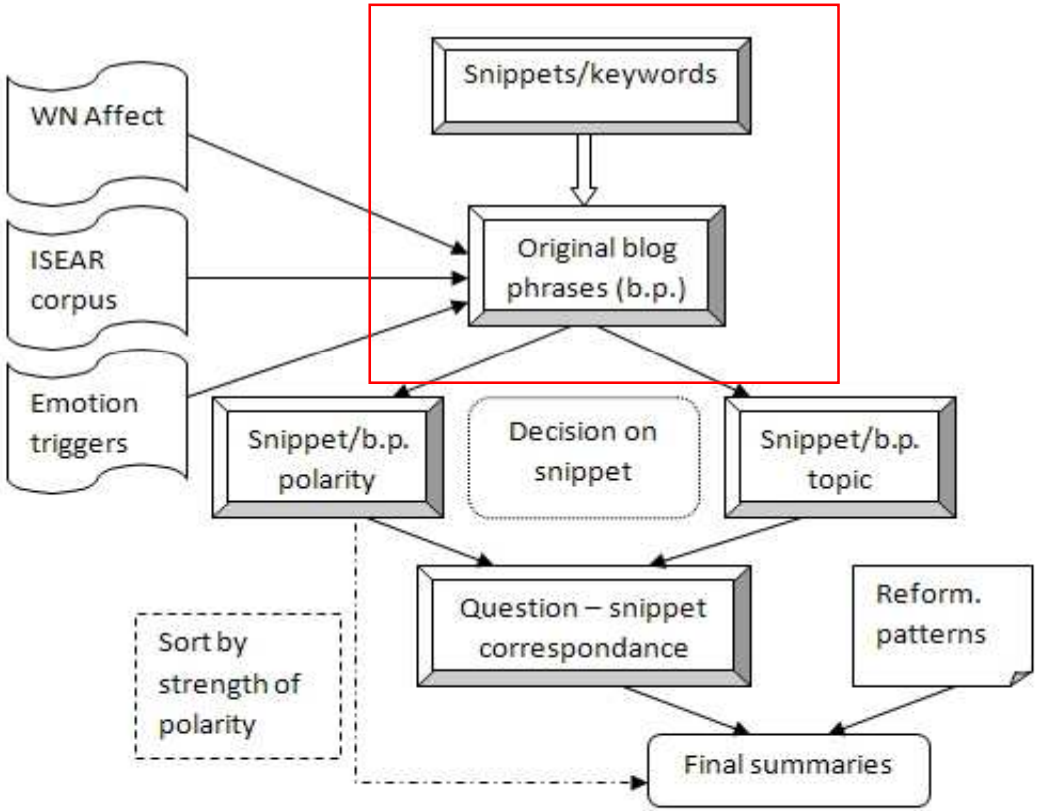
Answer retrieval

- 3 approaches, only 2 evaluated
  1. Using the provided answer snippets – snippet-driven approach
  2. Not using the provided snippets; including the blog answer candidate snippets – blog driven approach
  3. Using the provided answer snippets and employing anaphora resolution on original blogs

Snippet-driven approach

- **Blogs**
  - HTML tags removed; split into sentences
- **Using answer snippets provided**
  - Snippets sought in the original blogs
  - Those not literally contained -stemmed, stopwords removed
  - Computed similarity to potential sentences in the blogs with Pedersen's similarity package
  - Extract the most similar blog sentences, and their focus

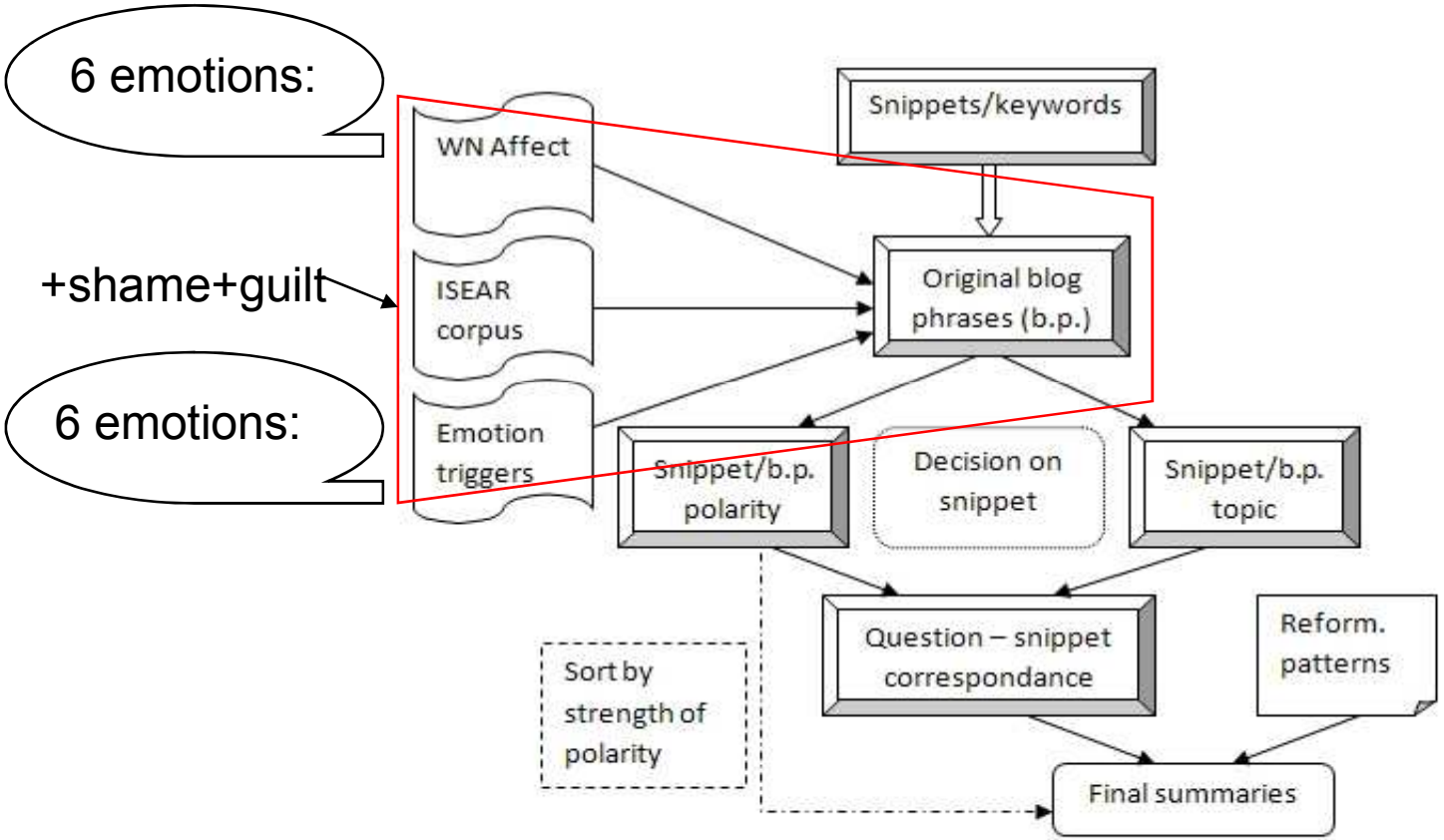
Snippet-driven approach



Snippet-driven approach

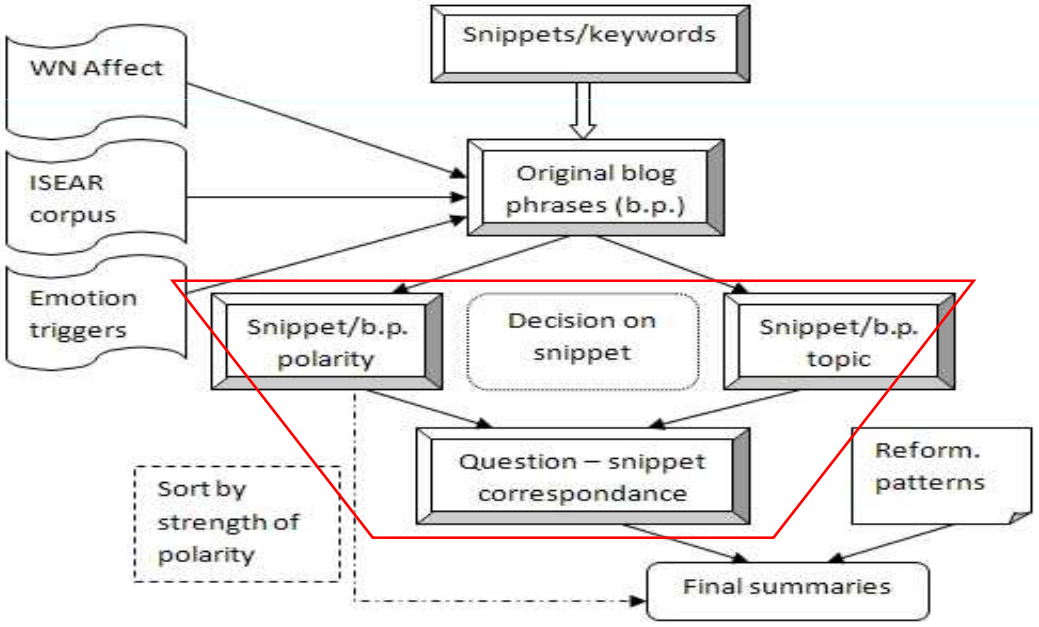
- **Eliminating "noise"**
  - Using Minipar and selecting only sentences with S and Pred
- **Determining the polarity of the snippet/blog phrase**
  - With Pedersen's Text Similarity Package, using the score with the terms in WN Affect, the ISEAR corpus and the emotion triggers
    - Summing up positive scores
    - Summing up negative scores
    - Which is the greater (no machine learning possibility)

Snippet-driven approach



Snippet-driven approach

- **Answering the questions**
  - By topic and polarity correspondance between the question and the retrieved snippets/blog phrases using the rules



Blog-phrase driven approach

Not using the answer snippet provided

- Eliminated the stopwords of the questions
- Determined the question focus&keywords
- Using the keywords and focus, determine blog phrases that could be the answer using similarity

Blog-phrase driven approach

- **Eliminating "noise"**
  - Using Minipar and selecting only sentences with S and Pred
- **Determining the polarity of the snippet/blog phrase**
  - With Pedersen's Text Similarity Package, using the score with the terms in WN Affect, the ISEAR corpus and the emotion triggers
- **Answering the questions**
  - By topic and polarity correspondance between the question and the retrieved snippets/blog phrases using the rules

Summary generation

- Using the question reformulation patterns and the retrieved answers;
- Tree-Tagger POS-Tagging to find 3rd pers. sing. and change them to 3rd pers. pl.;
- use replacement patterns(I/it etc)
- Snippet-driven: final summary
- Blog-driven: sorting the retrieved snippets in descending order, with respect to their polarity scores;included in summary those with highest scores, until reaching the imposed limit

Evaluation

- 1. summarizerID
- 2. Run type “manual”/ “automatic”
- 3. Use of answer snippets provided by NIST – “yes”/ ”no”
- 4. Average pyramid F-score (Beta=1), *averaged over 22 summaries
- 5. Grammaticality*
- 6. Non-redundancy*
- 7. Structure/Coherence *
- 8. Overall fluency/readability*
- 9. Overall responsiveness*

1	2	3	4	5	6	7	8	9
8	automatic	Yes	0.357	4.727	5.364	3.409	3.636	5.045
34	automatic	No	0.155	3.545	4.364	3.091	2.636	2.227

Table 2. *Evaluation results.*

0.534	7.545 (0.123)	7.63	3.591 (0.123)	5.318 (0.123)	5.409

Evaluation

- 1. summarizerID
- 2. Run type “manual”/ “automatic”
- 3. Use of answer snippets provided by NIST – “yes”/ ”no”
- 4. Average pyramid F-score (Beta=1), *averaged over 22 summaries
- 5. Grammaticality*
- 6. Non-redundancy*
- 7. Structure/Coherence *
- 8. Overall fluency/readability*
- 9. Overall responsiveness*

1	2	3	4	5	6	7	8	9
8	automatic	Yes	7	8	28	4	16	5
34	automatic	No	23	36	36	13	36	28

Table 3. *Classification results (overall comparison).*

Evaluation

- 1. summarizerID
- 2. Run type “manual”/ “automatic”
- 3. Use of answer snippets provided by NIST – “yes”/ ”no”
- 4. Average pyramid F-score (Beta=1), *averaged over 22 summaries
- 5. Grammaticality*
- 6. Non-redundancy*
- 7. Structure/Coherence *
- 8. Overall fluency/readability*
- 9. Overall responsiveness*

1	2	3	4	5	6	7	8	9
8	automatic	Yes	7	15	14	2	11	5
34	automatic	No	9	19	19	6	19	14

Table 4. *Classification results (comparison with systems using/not using answer snippets).*

Discussion

+ System performed well regarding Precision and Recall, the first run begin classified 7th among the 36 as F-measure
+ Structure and coherence 4/36 –reform. patterns
+ Overall responsiveness 5/36
+Second approach was well as F-measure – similarity/polarity/polarity strength
-- did not perform very well with respect of the non-redundancy criterion & grammaticality one

Conclusions

- With the participation in the TAC 2008 we could:

1. Test a general opinion mining system, working with different affect and opinion categories – worked well
2. Test the importance of the resources used and the relevance they have to this task – relevant resources
3. Test the relavance of polarity strength to the resultsand to computing the relevance of the retrieved text - positive
4. Test manners to generate coherence and grammaticality of text through patterns – evaluated well as coherence
5. Test a method of summarization based on polarity strength
6. Determine what is needed in order to build an MPQA system – a modified method from the classical QA systems

Future work

1. Employ a Textual Entailment system for redundancy detection
2. Check grammaticality
3. Develop alternative methods for retrieving the candidate answers, by query expansion, as for factual texts, but using affective and opinion vocabulary
4. Test how many of retrieved snippets were not included in summary due to polarity

Thank you!

Alexandra Balahur, Elena Lloret,
Andrés Montoyo, Manuel Palomar

Universitat d'Alacant
Universidad de Alicante

Department of
Software and
Computing Systems

gPLSI
Research Group of Language
Processing and Information
Systems