Overview of the TAC2013 Knowledge Base Population Evaluation: Sentiment Slot Filling

Margaret Mitchell
Microsoft Research
Redmond, WA, USA
memitc@microsoft.com

Abstract

This document provides an overview of the Text Analysis Conference 2013 Knowledge Base Population Sentiment Slot Filling (SSF) track. Sentiment Slot Filling was a new task added this year, and pushed the state of the art in sentiment by requiring teams to provide sentiment analysis in open domains across a variety of genres. The task focused on identifying the polarity of sentiment as well as sentiment holders and sentiment targets.

1 Introduction

One of the primary goals of Knowledge Base Population (KBP) at the Text Analysis Conference (TAC) is to develop technologies that can use unstructured text to populate knowledge bases about named entities. This year saw the introduction of the Sentiment Slot Filling (SSF) task, which aims to promote research into discovering sentiment expressed towards or by entities.

For this task, sentiment is defined as *a positive or negative emotion, evaluation, or judgement*. SSF therefore explores the sentiment triple:

<sentiment holder, sentiment, sentiment target>

Which we formalize as:

{query entity, sentiment slot} → filler entity

This task brings together two recent arcs in sentiment analysis, one focusing on recognizing holders, expressions, and targets (e.g., Yang and Cardie (2013)) and another looking at sentiment polarity targeted towards entities (e.g., Mitchell et al. (2013)). For TAC KBP 2013, entities may be a person (PER), organization (ORG), or a geopolitical entity (GPE).

The main challenges for this task therefore involve:

- Discovering entities that are holders and targets of sentiment
- Determining the polarity of the expressed sentiment
- Determining which entities across documents are the same as the query entity (cross-document co-reference resolution).

2 Task Definition

For Sentiment Slot Filling 2013, we are interested in collecting information on which entities hold sentiment towards another entity; which entities are recipients of sentiment from another entity; and what the polarity of the expressed sentiment is. This year, we limit our corpora to English texts.

Queries include a query entity and a sentiment slot that indicate both query polarity and directionality. Thus, depending on the sentiment slot, the query entity is either a sentiment holder or a sentiment target. Systems are required to return unique values for the remaining member of the triple: either sentiment targets or sentiment holders, depending on the sentiment slot.

For example, if the query specifies an entity with positive polarity towards X, systems must return distinct entities towards which the query entity holds a positive sentiment (the sentiment targets). If the query specifies an entity with negative polarity from X, systems must return distinct entities that hold negative sentiment towards the query entity (the sentiment holders). Possible answers therefore fill one of the following slots:

- **pos-towards:** query entity holds positive sentiment towards filler entity (likes, is hopeful about, etc.). In the triple <sentiment holder, sentiment, sentiment target>, the fillers are the *sentiment targets*.
- **pos-from:** query entity is target of positive sentiment from filler entity (is liked by, was

hoped for by, etc.). In the triple <sentiment holder, sentiment, sentiment target>, the fillers are the *sentiment holders*.

- **neg-towards:** query entity holds negative sentiment towards filler entity (dislikes, is skeptical about, etc.). In the triple <sentiment holder, sentiment, sentiment target>, the fillers are the *sentiment targets*.
- **neg-from:** query entity is target of negative sentiment from filler entity (is disliked by, etc.). In the triple <sentiment holder, sentiment, sentiment target>, the fillers are the *sentiment holders*.

Sentiment may be directed toward an entity based on direct evaluation of an entity (e.g., "Kentucky doesn't like Mitch McConnell") or may be directed to an entity based on actions that the entity took (e.g., "Kentucky doesn't like Mitch McConnell's stance on gun control"). In the current examples, given a query with {*Mitch McConnell*, **neg-from**}, the filler would be the holder of the sentiment, *Kentucky*.

2.1 Annotation guidelines

All four of the slots for Sentiment Slot Filling are name slots, meaning that they are required to be filled by the name of a person, organization, or geo-political entity (GPE):

- **Person Entities (PER)** - PERs are limited to individual humans. Groups of people (including families) are not valid person entities.
- **Organization Entities (ORG)** - ORGs are corporations, agencies, and other groups of people defined by an established organizational structure.
- **Geo-political Entities (GPE)** - Generally speaking, GPEs are composite entities comprised of a government, a physical location, and a population, with common types including countries, states, provinces, counties, cities, and towns.

The four Sentiment Slot Filling slots are list-valued, meaning that they can take multiple fillers. In the discussion forum and weblog data, post authors and bloggers may be used as query entities, or returned as filler entities, though they should only be used as query entities if they can be positively identified and thus either linked to the KB or marked as NIL. Complex uses of sarcasm were determined to be out of scope for this year.

2.2 Query format

Each query in the Sentiment Slot Filling task consists of a query ID, the name of the entity, a document (from the corpus) in which the name appears (to disambiguate the query in case there are multiple entities with the same name), the start and end offsets of the name as it appears in the document, its type (PER, ORG, or GPE), its KB node ID, and the sentiment slot to be filled (which specifies whether the query entity is a sentiment holder or a sentiment target, and the polarity of the sentiment held by or about them). An example query is:

```
<query id="SSF_ENG_002">
   <name>PhillyInquirer</name>
   <docid>eng-NG-31-141808-99662</docid>
   <beg>757</beg>
   <end>770</end>
   <enttype>ORG</enttype>
   <slot>pos-towards</slot>
   <nodeid>E0312533</nodeid>
</query>
```

2.3 Filler entities

As mentioned above, sentiment slot fillers are list-valued, where multiple fillers returned for the same query should refer to distinct individuals. It is not sufficient that slot filler entity strings be distinct; they must refer to distinct individuals. For example, if the query included {**Hillary Clinton**, *pos-towards*} (the sentiment holder is Hillary Clinton with positive sentiment towards the filler), and the system finds both "William Clinton" and "Bill Clinton" as potential fillers, just one of those strings should be returned. Similarly, entities should not be repeated as slot fillers for a query: Although it is possible that Hillary Clinton may feel *pos-towards* William Jefferson Clinton on many separate occasions, systems should only return one of these instances as a response.

The slot filler entity string returned by systems must be the most informative named mention of the entity in the document. For example, if "William Clinton" is the only named mention of the slot filler entity in the document, then it is acceptable to return that string as the slot filler; however, if "William Jefferson Clinton" is also in the document, then this more informative string should be returned.

To aid in this task, we provide standoff coreference chains and named entity tags for source documents using BBN's SERIF system (Ramshaw et al., 2001).

2.4 Provenance of query entity and filler entity

Systems are required to provide provenance information for both query entity and filler entity. Provenance is reported as start/end character offsets for the span of text which yielded the query entity or filler entity.

2.5 Justification

For each slot filler, systems must return sentences and clauses around the slot filler that provides justification for the extraction. The justification must contain at least one clause and at most two sentences. If two sentences are reported, they may be discontiguous. Further details on justification are provided in the TAC KBP 2013 Slot Filling guidelines.

3 Scoring and Assessment

The main difficulty with scoring slot filling systems that utilize such large corpora is that it is not feasible to prepare a comprehensive slot-filling answer key in advance; any manually-prepared key is likely to be incomplete. For this task, we approximate a comprehensive strategy by pooling the responses from all the systems and have human assessors judge the responses. To increase the chance of including answers which may be particularly difficult for a computer to find, LDC also prepares a manual key that is included in the pooled responses.

The slot filler in each non-NIL response is assessed as Correct, ineXact, or Wrong, as follows:

1. A response that contains more than two clauses/sentences in the justification is assessed as Wrong.

2. Otherwise, if the text spans defined by the offsets (+/- a few sentences on either side of each span) do not contain sufficient information to justify that the slot filler is correct, then the slot filler is also assessed as Wrong.

3. Otherwise, if the text spans justify the slot filler but the slot filler either includes only part of the correct answer or includes the correct answer plus extraneous material, the slot filler is assessed as ineXact. No credit is given for ineXact slot fillers.

4. Otherwise, if the text spans justify the slot filler and the slot filler string is exact, the slot filler is judged as Correct.

Team ID	Organizations
PRIS2013	Beijing University of Posts and Telecommunications
Columbia_NLP	Columbia University
CornPittMich	Cornell University, University of Pittsburgh

Table 1: Overview of participants for Sentiment Slot Filling, TAC KBP 2013

Two or more system responses for the same query entity and slot may have equivalent slot fillers (i.e., refer to the same entity); in this case, the system is given credit for only one response, and is penalized for all additional equivalent slot fillers. This is implemented by assigning each correct response to an equivalence class, and giving credit for only one member of each class.

Given these judgments, we can count:

- Correct = total number of correct equivalence classes in system responses
- System = total number of non-NIL system responses
- Reference = number of equivalence classes for all slots

Then:

- **Precision** = Correct / System
- **Recall** = Correct / Reference
- **F1** = 2*Precision*Recall / (Precision + Recall)

4 Participants and Systems

Participating teams are listed in Table 1. This task initially attracted 16 registered teams, out of which 3 teams submitted one or more runs. Two of the three teams noted that the lack of *cross-document co-reference* was a large barrier to entry; existing sentiment technology tends to work within-document, whereas TAC KBP requires responses for a given query entity across documents.

An overview of approaches for different runs from each team is shown in Table 2. Both the Columbia_NLP and CornPittMich teams followed

Team ID	Run ID	Description
PRIS2013	1	CRF results with web metadata and forum metadata
	2	CRF results with forum metadata
	3	CRF results
Columbia_NLP	1	Opinion detection
	2	Opinion detection, subject / object filter
	3	High confidence opinion detection
	4	High confidence opinion detection, subject / object filter
	5	High confidence opinion detection, subject / object filter, subjectivity assumed
CornPittMich	1	Pipeline with opinion extraction followed by holder / target extraction

Table 2: Overview of systems for Sentiment Slot Filling, TAC KBP 2013

a pipeline approach to identify holders/targets, subjective expressions, and sentiment polarity. The PRIS2013 team followed a relatively simpler pipeline, identifying holders/targets and using aggregate polarity over the whole sentence to determine the targeted sentiment.

In the Columbia_NLP system, a pipeline was used to first extract viable entity pairs, analyze the subjectivity of the text relating them, and then classify the polarity of the sentiment expressed. Similarly, in the PRIS2013 system, sentiment holders and targets were first identified, and then the polarity of the expressed sentiment was determined. In the CornPittMich system, subjective sentences and sentiment expressions were first identified, and then opinion holders/targets were associated to the identified sentiment.

A common approach among the teams was to use Conditional Random Fields (CRFs) (Lafferty et al., 2001) to identify sentiment holders and targets. The PRIS2013 team used two models based on CRFs, one to identify holders and one to identify targets. The CornPittMich team was a collaboration to combine two existing systems for fine-grained sentiment analysis, incorporating the CRF/ILP-based opinion analysis system of Yang and Cardie (Yang and Cardie, 2013) to identify subjective expressions, opinion targets, and opinion holders.

All three teams used the provided SERIF annotations for named entity recognition and coreference, and additionally brought in the Stanford CoreNLP[1] tools for dependency parsing (de Marneffe et al., 2006). All teams used some form of subjectivity or emotion lexicon, including those of (Wilson et al., 2005; Whissel, 1989; Stone et al., 1966).

Both the PRIS2013 team and the CornPittMich team also used the CoreNLP tools to provide POS-tagging, and the PRIS2013 additionally used the tools for further named entity annotations, nominal tagging, and chunking. Document retrieval techniques were different across all three teams.

5 Results

5.1 Scores

Table 3 lists results for the participating teams, with top scores in bold. From the PRIS2013 runs, we take the top 2000 system responses. The PRIS2013 produced the most reliable results overall, reaching an F-Score of 13.15%. The CornPittMich team had best system precision at 10.00%. Relaxing the justification requirements in Table 4, the PRIS2013 achieved the highest scores overall.

Table 5 illustrates the number of correct slots in the top systems from all teams, for Newswire (NEWS), Web Text (WEB), and Discussion Fora (FORA). Across teams, very few correct responses were drawn from the Web data. Discussion fora provided the richest source of correct slot fillers for this task. There is also a slight trend for the *towards* slots to come from NEWS sources, and the *from* slots to come from FORA sources. However, responses are not reliable enough at this point to come to firm conclusions about what corpora and techniques are best for this task.

A key difference between the systems is that while PRIS2013 developed a broad approach, using similar models for Sentiment Slot Filling and regular Slot Filling, both the CornPittMich team and the Columbia_NLP team[2] focused their devel-

[1] http://nlp.stanford.edu/downloads/corenlp.shtml

[2] The Columbia_NLP team's original submission returned

Team ID	Run	Prec.	Rec.	F1
PRIS2013	1	9.15	20.24	12.60
	2	9.55	**21.13**	**13.15**
	3	5.50	12.17	7.58
Columbia_NLP	1	1.81	0.44	0.71
	2	1.83	0.44	0.71
	3	1.68	0.22	0.39
	4	1.72	0.22	0.39
	5	1.67	1.00	1.25
CornPittMich	1	**10.00**	0.77	1.44
LDC		70.01	75.66	72.73

Table 3: Official scores for Sentiment Slot Filling: Precision (Prec.), Recall (Rec.) and F-Score (F1) in %. The LDC score corresponds to the output created by the LDC experts.

	IGNOREOFFSETS			ANYDOC		
Team ID	**P**	**R**	**F1**	**P**	**R**	**F1**
PRIS2013	10.1	22.4	13.9	11.5	25.5	15.9
Columbia_NLP	1.9	1.1	1.4	2.6	1.6	2.0
CornPittMich	8.6	0.7	1.2	10.0	0.8	1.44

Table 4: Results for Sentiment Slot Filling, best team runs, ignoring justification: Precision (P), Recall (R) and F-Score (F1) in %. In the IGNOREOFFSETS configuration, justifications are considered correct if the correct document is reported. In the ANYDOC configuration, justifications are completely ignored, and fillers are marked as correct solely based on string matching with gold fillers. The LDC score corresponds to the output created by the LDC experts.

	Data source			
Slot	NEWS	WEB	FORA	**Total**
pos-towards	14	0	3	17
neg-towards	15	1	10	26
pos-from	11	0	72	83
neg-from	8	1	72	81
Total	48	2	157	207

Table 5: Number of correct slots in the top systems from all teams, for Newswire (NEWS), Web Text (WEB), and Discussion Fora (FORA).

(a)

Count	Assessment
4124	Wrong
407	Correct
126	Inexact-Long
282	Inexact-Short
221	Ignore

(b)

Count	Assessment
3947	Wrong
965	Correct
27	Inexact
221	Ignore

Table 6: Pooled assessment results for relation justification (a) and slot filler correctness with respect to justification (b).

opment on finding fillers *within the same document* as the query entity. It is beneficial in this context to look at both **precision** and **recall**. While *recall* measures how well systems performed at retrieving correct answers across documents, *precision* measures how well systems performed on the responses they did make; focusing on sentiment within the same document as a query entity has the general effect of trading higher precision for lower recall, and we see this trend here.

5.2 Error Analysis

Assessment results on justification offsets for the pooled responses are shown in Table 6 (a), and assessment results on the correctness of the pooled slot fillers are shown in Table 6 (b). Most justifications were assessed to be Wrong (4124). When a justification was inexact, it was usually too short. Relatively few slot fillers were assessed to be inexact. Many responses included offsets that were too long for assessors to read, resulting in an **Ignore** assessment and removed from scoring.

A qualitative analysis of the errors across systems suggests that although systems were retrieving entities, query and entity offsets were often incorrect. Justifications were often inexact, without enough text; or else too long for assessors to read. The <holder, target> relationship was occasionally reversed by systems, which affected their output responses. Perhaps most significantly, it is clear that detecting the correct polarity of targeted sentiment remains a challenging task.

6 Concluding Remarks

This year marks the first Sentiment Slot Filling task for TAC KBP. A primary challenge of this task is to find slot fillers for a query within the KBP corpora. Query entities were identified with a KB node ID (if available) and single document identifiers; teams therefore had to determine whether a given query entity was the same as an entity found in another document (cross-document co-reference) and/or determine whether an entity in a document is the same as the one in the knowl-

fillers solely within-document; preliminary results suggest that their F-score greatly improves when adapted to make use of the full corpus.

edge base (entity linking). This is an extremely challenging task in its own right, and two of the three submitted systems were originally developed to retrieve slot fillers within the same document as the query entity, without addressing the cross-document difficulties.

These initial results are promising, but suggest that teams and TAC KBP alike should focus on ways to better connect query entities to references throughout the documents. Looking forward to next year, we may achieve further gains by limiting queries and system responses to be within the same document; using a small subset of the designated KBP documents; or else providing better tools for cross document co-reference resolution and entity linking throughout the corpora.

Acknowledgments

This task was made possible thanks to the input and expertise of Boyan Onyshkevych, Theresa Wilson, Ben Van Durme, Mihai Surdenau, Hoa Trang Dang, Joe Ellis, Kira Griffit, Stephanie Strassel, and the rest of the KBP organizers. We thank Hoa Dang for implementing the SSF scorers, John Frank and Elizabeth Boschee for running BBN's SERIF, and Mihai Surdenau for assistance writing guidelines. We also gratefully acknowledge the participants, particularly Sara Rosenthal and Claire Cardie, for providing useful feedback.

References

M.-C. de Marneffe, B. MacCartney, and C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. *Proceedings of the Fifth International Conference on Language Resources and Evaluation.*

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML-2001.*

M. Mitchell, J. Aguilar, T. Wilson, and B. Van Durme. 2013. Open domain targeted sentiment. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

L. Ramshaw, E. Boschee, S. Bratus, S. Miller, R. Stone, R. Weischedel, and A. Zamanian. 2001. Experiments in multi-modal automatic content extraction. *Proceedings of the First International Conference on Human Language Technology Research.*

P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press, Cambridge.

C. M. Whissel. 1989. The dictionary of affect in language. In *Emotion: theory research and experience*, volume 4, London. Academic Press.

T. Wilson, J. Wiebe, and P. Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.*

B. Yang and C. Cardie. 2013. Joint inference for fine-grained opinion extraction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.*