Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Using Frame Semantics for Knowledge Extraction from Twitter

Anders Søgaard, Barbara Plank, and Hector Martinez Alonso
Center for Language Technology, University of Copenhagen, Denmark
soegaard@hum.ku.dk

Abstract

Knowledge bases have the potential to advance artificial intelligence, but often suffer from recall problems, i.e., lack of knowledge of new entities and relations. On the contrary, social media such as Twitter provide abundance of data, in a timely manner: information spreads at an incredible pace and is posted long before it makes it into more commonly used resources for knowledge extraction. In this paper we address the question whether we can exploit social media to extract new facts, which may at first seem like finding needles in haystacks. We collect tweets about 60 entities in Freebase and compare four methods to extract binary relation candidates, based on syntactic and semantic parsing and simple mechanism for factuality scoring. The extracted facts are manually evaluated in terms of their correctness and relevance for search. We show that moving from bottom-up syntactic or semantic dependency parsing formalisms to top-down frame-semantic processing improves the robustness of knowledge extraction, producing more intelligible fact candidates of better quality. In order to evaluate the quality of frame semantic parsing on Twitter intrinsically, we make a multiply frame-annotated dataset of tweets publicly available.

Knowledge extraction has primarily focused on mining Wikipedia and newswire data. For this reason, knowledge bases used for search such as Freebase suffer from low recall, only covering certain entity types, and only certain facts about those entities. Freebase, for example, contains the fact that the Walt Disney Company is a production company, but not that it, for instance, owns Marvel.

A common problem in knowledge extraction is what is known as the *reporting bias* (Gordon and van Durme 2013), i.e., the fact that a lot of common knowledge is never made explicit. Social media platforms like Twitter have the potential to fill some of that gap, since they offer very different facts than what can be found in Wikipedia. People may tweet an obvious fact to inform their friends what they just realized, as a means of sarcasm, or simply to kill time. Finally, Twitter is a platform that potentially allows us to harvest facts in almost real time. E.g., a company may buy up another company and tweet about it, long before the fact makes it into Wikipedia or Freebase.

On the other hand, extracting useful facts from Twitter is a hard problem. Tweets often contain opinionated non-factual text, automated posts from third-party websites, and/or temporary facts that are irrelevant to search. True facts seem like needles in haystacks, but, on the other hand, the haystacks are plentiful on Twitter.

There is also another reason that knowledge extraction from Twitter is hard. Most approaches to knowledge extraction rely on syntactico-semantic processing, and state-of-the-art parsing models fair badly on Twitter data (Foster et al. 2011), to the extent that it is prohibitive for downstream applications such as knowledge extraction.

In this paper, we show that top-down frame semantic parsing is more robust to the domain shift from newswire to Twitter than other syntactico-semantic formalisms, and that this leads to more robust knowledge extraction. In particular, while syntactic and semantic dependency parsing models induced from newswire exhibit dramatic drops when applied to Twitter data, frame semantic parsing models seem to perform almost the same across domains.

Our Approach We select 60 entities in Freebase distributed equally across persons, locations and organizations (see Table 1), and extract 70k tweets mentioning at least one of these entities. The data was collected during the summer 2014. We part of speech (POS) tag these tweets and pass the augmented tweets on to four different extraction models: a syntactic dependency parser, a semantic role labeler, a frame semantic parser, and a rule-based off-the-shelf (REVERB) open information extraction system (Fader, Soderland, and Etzioni 2011). For all systems, except REVERB, we apply the same heuristics to filter out relevant facts and rank them in terms of factuality using sentiment analysis. We evaluate facts in terms of their wellformedness, their correctness, and their relevance. We also ask subjects to rate triples in terms of opinionatedness for the sake of error analysis. Finally, we check the extracted facts for novelty against Freebase.

Frame Semantic Parsing

Frame semantic parsing is the task of assigning frames (Fillmore 1982) to text. Frames combine word sense disambiguation and semantic role labeling. The go-to ressource

PERSON		LOCATION		ORGANIZATION	
Kurt Cobain	Andy Warhol	Stonehenge	Kakadu Natl. Park	Opus Dei	EU Parliament
Janis Joplin	Jean-Michel Basquiat	Ground Zero	Ningaloo Reef	Oddfellow	Monsanto
Brian Eno	Paul Cezanne	Yellowstone	Himeji-jo	Wikileaks	PETA
Miles Davis	Jeff Koons	Grand Canyon	Polonnaruwa	MI6	NRA
Amy Winehouse	Frida Kahlo	Mount Everest	Hampi	Freemasons	Pixar
Peter Greenaway	James Dean	Mount Fuji	Angkor Wat	Greenpeace	Disney
Rainer W. Fassbinder	Lana Turner	Acropolis	Ha Long Bay	Vanderbilt	Microsoft
Kenneth Anger	Rita Hayworth	Clonmacnoise	Skogskyrkogården	NSA	General Motors
Man Ray	Elizabeth Taylor	Pompeii	Christiansø	FBI	Procter&Gamble
David Lynch	Marlon Brando	Ayers Rock	Lunenburg	CIA	Walmart

Table 1: Entity seeds used in our experiments. Shaded cells not discovered in Twitter

is FrameNet, a taxonomy of manually identified general-purpose frames for English. Each frame comes with a set of lemmas (with POS), called lexical units, whose associated word forms can potentially trigger the given frame. These word forms are called *targets*. Each frame definition also includes core and peripheral roles, such as participants and attributes, called the *arguments* of the frames. Frame semantic parsing can be thought of as the problem of deciding whether a word form triggers a frame in a specific context, what frame that would be, and what arguments are expressed by what parts of the context (if at all).

FrameNet also includes a set of 139K lexicographic exemplar sentences. This amounts to 3.1M words. The examples primarily come from the British National Corpus. Another resource is the data from the SemEval 2007 shared task on semantic role labeling.[1] This data comes from bureaucratic texts and newswire and consists of 43.3K training sentences, as well as development and test data. The SEMAFOR system used in our experiments below was trained on the FrameNet exemplar sentences and the SemEval 2007 data. All the data is available on the FrameNet websites.[2]

This paper argues that frame semantics is particularly appropriate for the semantic analysis of text on Twitter. Since the vast majority of annotated parsing ressources come from newswire and related domains, the quality of linguistic analysis on Twitter typically depends on the cross-domain robustness of our models. One of the key differences between newswire and Twitter is the differences in how people use determiners, capitalization, function words and punctuation. The intuition behind this paper is that while these are important signals for parsing models predicting complete structures (connecting all words in a sentence or a post), these tokens are less important for frame semantic parsing, producing only partial structures.

In support of this argument, we present downstream results for knowledge extraction using frame semantics, but we would also like to provide intrinsic evaluations of frame semantic parsing models induced from newswire on Twitter data. Since such data is not available, we present our own frame semantic annotations of Twitter below.

[1] http://nlp.cs.swarthmore.edu/semeval/tasks/task19/

[2] https://framenet.icsi.berkeley.edu/

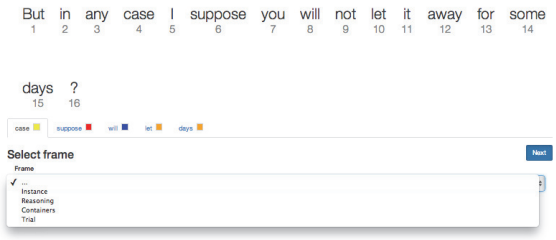

Figure 1: Browser-based annotation interface.

Annotations

We present Twitter data annotated with frames from FrameNet 1.5. Rather than annotating raw text from scratch, we chose to annotate the development and evaluation splits of an annotated Twitter corpus used in previous studies (Ritter et al. 2011; Derczynski et al. 2013).[3] The splits are those provided by (Derczynski et al. 2013).

We created a software for frame semantic annotation of POS tagged text with a web browser interface. The software pre-annotates the text with possible frames that annotators select from drop-down menus. The arguments are identified by the index of their head word. Annotators were asked to focus on core roles, and only provide peripheral roles when they played prominent roles in the sentence. See Figure 1 for a screen dump.

All tweets in the corpus were annotated by the same three annotators. Rather than adjudicating between these annotators to arrive at a single gold standard, we compute figures for all three annotations in our experiments below, motivated by recent arguments that adjudication biases evaluation (Plank, Hovy, and Søgaard 2014). In order to make tables easier to read, we provide averages over the three annotators when we report our results below.

Note that one difference between the exemplar sentences and the SemEval 2007 data on the one hand, and our annotations on the other hand, is that frames and their arguments may potentially go across sentence boundaries. Here is a nice example:

[3] https://github.com/aritter/twitter_nlp

(1) Funniest thing I_{Hearer} **heard**$_{Hear}$ this week. Wingo telling me and wood. "I'm scared$_{Message}$."

Here, *heard* is the target evoking the frame, in this case the frame HEAR. This frame has two explicit arguments, a hearer and a message. The speaker is not explicit in this case.

The inter-annotator agreements were high between our annotators. The target identification F_1 scores ranged between 91.76 and 95.60, and the frame identification F_1 scores between 83.05 and 85.44. See §4 for these metrics.

Parser

The SEMAFOR system (Das et al. 2010) uses a small set of rules to identify potential targets. Targets must appear in the lexicon or the annotated training data. The system also disregards all prepositions, which is a potential source of error (see §4). The frame identification step predicts a frame given a word form in context with the triggering lemma as a hidden variable, using logistic regression. Note that (Das et al. 2010) make a strong independency assumption, predicting frames independently of each other. Given a list of identified frames, SEMAFOR identifies arguments by another log-linear model, given a sentence, a target, a frame, an argument and a candidate span. SEMAFOR subsequently enforces that arguments of the same frame cannot have overlapping spans, by reranking. This is a potential source of error in our case, since our annotators were allowed to assign multiple roles (of the same frame) to the same tokens. We use an in-house Twitter-adapted POS tagger.

Results

We evaluate the performance of SEMAFOR on our annotated Twitter data using various metrics, previously introduced in (Das et al. 2010). All scores are averages over the three annotations. TARGET IDENTIFICATION is the F_1 score balancing how many frame-evoking tokens the system predicts to be triggers, and how often the system is right in predicting a token to envoke a frame. FRAME IDENTIFICATION is basically the labeled F_1 score, which in addition to TARGET IDENTIFICATION requires also getting the frame label right. Our ARGUMENT IDENTIFICATION F_1 metric is a little different from the metric in (Das et al. 2010), in measuring the system's ability to predict arguments irrespectively of getting the frame label right. The reason for this choice is that we asked annotators to focus mainly on core roles, and only supply peripheral roles if they felt they were important. The exact frame matching F_1 scores, including exactly matching all arguments, range between 20.26 and 25.96 on RITTER-EVAL across our three annotators.

Note that these results are surprisingly comparable to previously reported results on newswire-like text; cf. line 4. (Das et al. 2010) obtain a target identification F_1 score of 79.21, a frame identification F_1 score of 61.44, and an exact frame matching score of 46.49. While results are obviously not directly comparable, this suggests, exact frame matching aside, that there is little or no cross-domain loss for frame semantics. The relative robustness across domains of frame semantic parsing is in sharp contrast to syntactic parsing (Foster et al. 2011). However, compared to inter-annotator agreement (IAA; line 1), there is still room for improvement.

Knowledge Extraction

Data

Our dataset consists of tweets containing one of a set of 60 pre-defined entities (20 persons, 20 locations, and 20 organizations). We queried Twitter over the course of a few days to extract tweets containing any of those entities. This resulted in a corpus of 70,000 tweets. Three entities never occurred in the search results. The tweets were then POS-tagged with a Twitter-adapted POS tagger (Derczynski et al. 2013)[4] and passed on to four different extraction systems, described below.

Extraction Systems

Dependency Parsing Our dependency parser is the graph-based parser of (Bohnet 2010), available as part of MATE-TOOLS.[5] We use default parameters. In extraction, we only consider verbs with two outgoing dependencies labeled as subject, object or predicate. These can be long-distance dependencies, and we therefore expect a graph-based parser to perform better than a transition-based parser on this task (McDonald and Nivre 2007).

Semantic Role Labeling We use the semantic role labeler described in (Björkelund et al. 2010), which is also distributed as part of MATE-TOOLS. In extraction, we only consider ARG0, ARG1 and ARG2 relations.

Frame Semantic Parsing We use *Semafor* (see §*Frame semantics*).

Rule-based Extraction. As a baseline and sanity check, we also ran the rule-based relation extraction system REVERB out of the box (Fader, Soderland, and Etzioni 2011). This system identifies relations that satisfy a series of syntactic and lexical constraints, and assigns as arguments the surrounding noun phrases. The relations are scored by a confidence function that uses logistic regression trained on a set of relations extracted from Wikipedia and labelled as correct or incorrect.

For all systems, we disregard triples with arguments headed by stop words. In the extraction step, we also disregard modifiers, keeping only the head word, with the exception of multi-word proper nouns referring to named entities.

Extraction

We apply various heuristics to whittle down the stream of extracted candidates. We require (i) that the first argument of the extracted triple is one of the 60 target entities, (ii) that the

[4]This tagger is slightly different from the one used in the intrinsic evaluation, e.g., it is also trained on RITTER-TEST. Therefore, we could not use this tagger in the frame semantic parsing experiments.

[5]https://code.google.com/p/mate-tools/

DATA	SYSTEM	TARGET IDENT (F_1)	FRAME IDENT (F_1)	ARG IDENT (F_1)
RITTER-TEST	IAA	95.3	84.5	78.1
	SEMAFOR-O	78.4	58.3	66.5
	SEMAFOR-W	81.9	62.1	67.5
SEMEVAL-07	SEMAFOR-O	79.2	61.4	46.5

Table 2: SEMAFOR results on Twitter data w/o gold POS tags, compared to results on newswire (SEMEVAL-07). Note the small drop from line 4 (newswire) to line 2 (Twitter). Same metrics as (Das et al. 2010). FRAME IDENTIFICATION and ARGUMENT IDENTIFICATION are exact matching scores.

most frequent sense of the verb has the super sense *stative*, *possession*, *competition* or *creation* according to Princeton WordNet, and (iii) that none of the arguments are stop words (closed class items). We then extract triples made up of a verb and the head words of the two arguments, e.g., *compose(Brian_Eno, music)*, and evaluate whether they express true and relevant facts (here: BRIAN ENO COMPOSES MUSIC).

The triples are then ranked as follows: Given a sample of tweets labeled as positive, negative, and neutral, we collapse the positive and negative classes into one class and train a logistic regression classifier to score tweets with respect to neutrality (which we use as a proxy for factuality). We use the labeled data from the SemEval 2013 shared task on sentiment analysis on Twitter and the same feature model as the second best system in the shared task.

We present examples of extracted facts by the different systems, with different verb types, and their predicted sentiment scores in Table 3. The sentiment scores (NEUTRALITY) are the confidence estimates of the binary logistic regression model that the tweet is neutral with respect to polarity. Note that facts where the target word is the second argument, are discarded in the extraction step.

Evaluation

We had three professional annotators (cf. Table 4) annotate the top 100 fact candidates from each system. The facts were rated as INTELIGIBLE, TRUE, OPINIONATED and RELEVANT. If a fact was said to be uninteligible, annotators were not asked to annotate whether it was true, opinionated or relevant. TRUE and RELEVANT almost correspond to *accuracy* and *usefulness* in (Mayfield et al. 2012), but note that we ignore recall, since we only consider Twitter a supplementary resource for knowledge extraction.

Results

The results are presented in Table 4. With syntactic parsing, about 63% of the extracted facts are inteligible. Out of these, 83% were judged to be true. 12% were judged as opinionated, and 87% as relevant for knowledge extraction. This means that in total 53/100 facts were true and inteligible facts that could be put in a knowledge base. For semantic role labeling, results are a little better; and for frame semantics, even better, with 58/100 facts being true *and* inteligible.

The professional annotators' ranking of the systems is clear across all metrics: FRAMES$\prec$SRL$\prec$SYNTAX$\prec$REVERB. The difference in inteligibility going from semantic-role labeling to frame semantic parsing *is* statistically significant ($p \sim 0.01$, computed using Wilcoxon's test). Note that the reduction in uninteligible facts is more than 20%.

Generally, our sentiment analysis models succesfully filters out most of the opinionated facts. This classifier, as mentioned above, is a simple logistic regression classifier trained on a bag-of-words model. We expect there to be considerable room for improvement, introducing more sophisticated sentiment analysis models.

We note that while frame semantic parsing is more robust than the other parsing models, it seems to extract (proportionally) fewer true facts. We currently do not have an explanation for this observation.

Freebase Coverage

We also tested whether the extracted facts were already in Freebase. Obviously, the names of the relations we identify are not those of Freebase relations, but it was easy, on the other hand, to see how many entity pairs in our triples were paired in Freebase. While the 60 target entities were all in Freebase, less than 1% of the entity pairs we found with dependency parsing were in Freebase, whereas 41% of the second argument entities were in Freebase (10% were in Princeton WordNet).

Error Analysis

POS Tagging and Parsing Errors

For error analysis, we asked one of our annotators to also judge whether unintelligible triples were likely to be due to POS tagging errors. In about half of the cases, the annotator found this to be the likely cause of error. The predicted head nouns included *gets, teary, could, available*, and *must*. The remaining unintelligible facts seemed to be either due to parsing errors such as DAVID LYNCH GETS GAME, where *game* was falsely predicted to be an object of *gets*, or due to metaphor, such as in NSA IS BAZAR.

Extraction Errors

Our extraction pipeline makes a number of controversial assumptions at present. First, we do not try to disambiguate the argument words in the extracted triples. The fact POMPEII IS A MOVIE is true, but obviously not of Pompeii, the location. Only extracting head words as arguments also introduces potential information loss. Consider, for example, the sentence:

SYSTEM	VERB SENSE	EXAMPLE	NEUTRALITY
SYNTAX	v.stative	publish(Wikileaks, mails)	0.95
	v.creation	paint(Andy Warhol, Debbie Harry)	0.91
	v.change	start(David Lynch, transcendental meditation)	0.97
	v.stative	is(NSA, enemy)	0.05
	v.social	joins(Al Qaeda, Greenpeace)	0.02
SRL	v.possession	sell(Wallmart, pillows)	0.95
	v.creation	produce(Brian Eno, "The Unforgettable Fire")	0.95
	v.possession	finance(CIA, Afghan Muhajideen)	1.00
	v.motion	send(Greenpeace,ship)	0.89
	v.communication	interrupt(Pharrell, Jeff Koons)	0.20

Table 3: Example candidate facts. Gray-shaded parts violate our extraction heuristics. *: From *Pompeii is the best song.*

	SELECTION	INTELIGIBLE	TRUE	OPINIONATED	RELEVANT
SYNTAX	All	0.63	0.53	0.08	0.55
	Inteligible	-	0.83	0.12	0.87
SRL	All	0.71	0.54	0.12	0.58
	Inteligible	-	0.76	0.17	0.82
FRAMES	All	**0.77**	**0.58**	0.14	**0.66**
	Inteligible	-	0.76	0.18	0.86
REVERB	All	0.64	0.39	**0.23**	0.42
	Inteligible	-	0.61	0.36	0.65

Table 4: Human judgments.

(1) Pompeii is the nicest place in Italy.

Here we can only extract the fact that POMPEII IS A PLACE, not POMPEII IS IN ITALY.

Judgment Noise

The human judgments are obviously not in perfect agreement. Some annotators may not know that POMPEII IS A MOVIE, for example. Some triples may be inherently ambiguous, such as POMPEII IS A DISASTER, which can be a true fact about a location, or an opinionated statement about a movie or a song.

Related Work

There has to the best of our knowledge been no previous work on relation extraction from Twitter, but a fair amount of work on event extraction exists. (Benson, Haghighi, and Barzilay 2011) use distant supervision in the form of a concert calendar to extract (concert) events from Twitter. (Becker et al. 2012) query Twitter and other social media platforms to automatically build event descriptions. In the same vein, (Balahur and Tanev 2013) discuss how to find tweets relevant to real-world events from the news.

Since tweets often describe events (here and now), the focus on event extraction rather than knowledge extraction is unsurprising. We believe, however, that our results in this paper indicate that Twitter is potentially an invaluable resource for knowledge extraction.

Conclusion

In this paper, we investigate knowledge extraction from Twitter. We use four different approaches to extracting facts about 60 entities in Freebase, and evaluate them along several dimensions. We find that given correct syntactic analysis we can extract true and relevant knowledge that is *not* already in Freebase with high precision. However, for most systems about two out of three triples were judged unintelligible, due to poor POS tagging and dependency parsing. We show that frame semantics provides more robust results, reducing more than 20% of the errors due to unintelligibility.

Acknowledgements

We would like to thank the anonymous reviewers for valuable comments and feedback. This research is funded by the ERC Starting Grant LOWLANDS No. 313695, as well as by the Danish Research Council Grant No. 1319-00123.

References

Balahur, A., and Tanev, H. 2013. Detecing event-related links and sentiments from social media texts. In *ACL*.

Becker, H.; Iter, D.; Naaman, M.; and Gravano, L. 2012. Identifying content for planned events across social media sites. In *WSDM*.

Benson, E.; Haghighi, A.; and Barzilay, R. 2011. Event discovery in social media feeds. In *ACL*.

Björkelund, A.; Bohnet, B.; Hafdell, L.; and Nugues, P. 2010. A high-performance syntactic and semantic dependency parser. In *COLING*.

Bohnet, B. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *COLING*.

Das, D.; Schneider, N.; Chen, D.; and Smith, N. 2010. Probalistic frame-semantic parsing. In *NAACL*.

Derczynski, L.; Ritter, A.; Clark, S.; and Bontcheva, K. 2013. Twitter part-of-speech tagging for all: overcoming sparse and noisy data. In *RANLP*.

Fader, A.; Soderland, S.; and Etzioni, O. 2011. Identifying relations for open information extraction. In *EMNLP*.

Fillmore, C. 1982. Frame semantics. In *Linguistics in the morning calm*. Hanshin.

Foster, J.; Cetinoglu, O.; Wagner, J.; Roux, J. L.; Nivre, J.; Hogan, D.; and van Genabith, J. 2011. From news to comments: Resources and benchmarks for parsing the language of Web 2.0. In *IJCNLP*.

Gordon, J., and van Durme, B. 2013. Reporting bias and knowledge extraction. In *The 3rd Workshop on Knowledge Extraction, CIKM*.

Mayfield, J.; Dorr, B.; Finin, T.; Oard, D.; and Piatko, C. 2012. Knowledge base evaluation for semantic knowledge discovery. In *Symposium on Semantic Knowledge Discovery, Organization and Use*.

McDonald, R., and Nivre, J. 2007. Characterizing the errors of data-driven dependency parsers. In *EMNLP-CoNLL*.

Plank, B.; Hovy, D.; and Søgaard, A. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *EACL*.

Ritter, A.; Clark, S.; Etzioni, O.; et al. 2011. Named entity recognition in tweets: an experimental study. In *EMNLP*.