The Holy Grail of Sense Definition: Creating a Sense-Disambiguated Corpus from Scratch

Anna Rumshisky
Marc Verhagen
Jessica Moszkowicz

September 18, 2009
GL2009 – Pisa, Italy

Talk Outline

- Problem of Sense Definition
- An Empirical Solution?
- Case Study
- Evaluation
- Constructing a Full Resource: Issues and Discussion

Problem of Sense Definition

- Establishing a set of senses is a task that is notoriously difficult to formalize
  - In lexicography, "lumping and splitting" senses during dictionary construction is a well known problem
  - Within lexical semantics, there has been little consent on theoretical criteria for sense definition
  - Impossible to create a consistent, task-independent inventory of senses

Standardized Evaluation of WSD and WSI Systems?

- Within computational community, a sustained effort to create a standardized framework for training and testing word sense disambiguation (WSD) and induction (WSI) systems
  - SenseEval competitions (2001, 2004, 2007)
  - Shared SRL tasks at the CoNNL conference (2004, 2005)
- Creating a gold standard in which each occurrence of the target word is marked with the appropriate sense from a sense inventory.

Sense Inventories

- Taken out of MRDs or lexical databases
    - WordNet, Roget's thesaurus, LDOCE
- Constructed or adapted from an existing resource in pre-annotation stage
    - PropBank, OntoNotes

Sense Inventories

- Choice of sense inventory determines the quality of the annotated data
  - e.g. SemCor (Landes et al, 1998) uses WordNet synsets, with senses that are too fine-grained and often poorly distinguished
- Efforts to create coarser-grained inventories out of existing resources
  - Navigli (2006), Hovy et al (2006), Palmer et al. (2007), Snow et al. (2007)

Creating a Sense Inventory

- Numerous attempts to formalize the procedure for creating a sense inventory
  - FrameNet (Ruppenhofer et al, 2006)
  - Corpus Pattern Analysis (Hanks & Pustejovsky, 2005):
  - PropBank (Palmer et al., 2005)
  - OntoNotes (Hovy et al., 2006)
- Each involves somewhat different approaches to corpus analysis done to create or modify sense inventories

Empirical Solution to the Problem of Sense Definition

- Create both a sense inventory and an annotated corpus at the same time
- Using native speaker, non-expert annotators
- Very cheap and very fast

Amazon's "Mechanical Turk"

- Introduced by Amazon as "artificial artificial intelligence"
  - "HITs": human intelligence taks, hard to do automatically, very easy for people
- Used successfully to create annotated data for a number of NLP tasks (Snow et al, 2008), robust evaluation for machine translation systems (Callison-Burch, 2009).
  - Complex annotation split into smaller steps
  - Each step farmed out to non-expert annotators ("Turkers")

Annotation Task

- A task for Turkers designed to imitate the process of creating clusters of examples used in Corpus Pattern Analysis
- In CPA, a lexicographer sorts a set of instances for a given target word into clusters according to sense-defining syntactic and semantic patterns

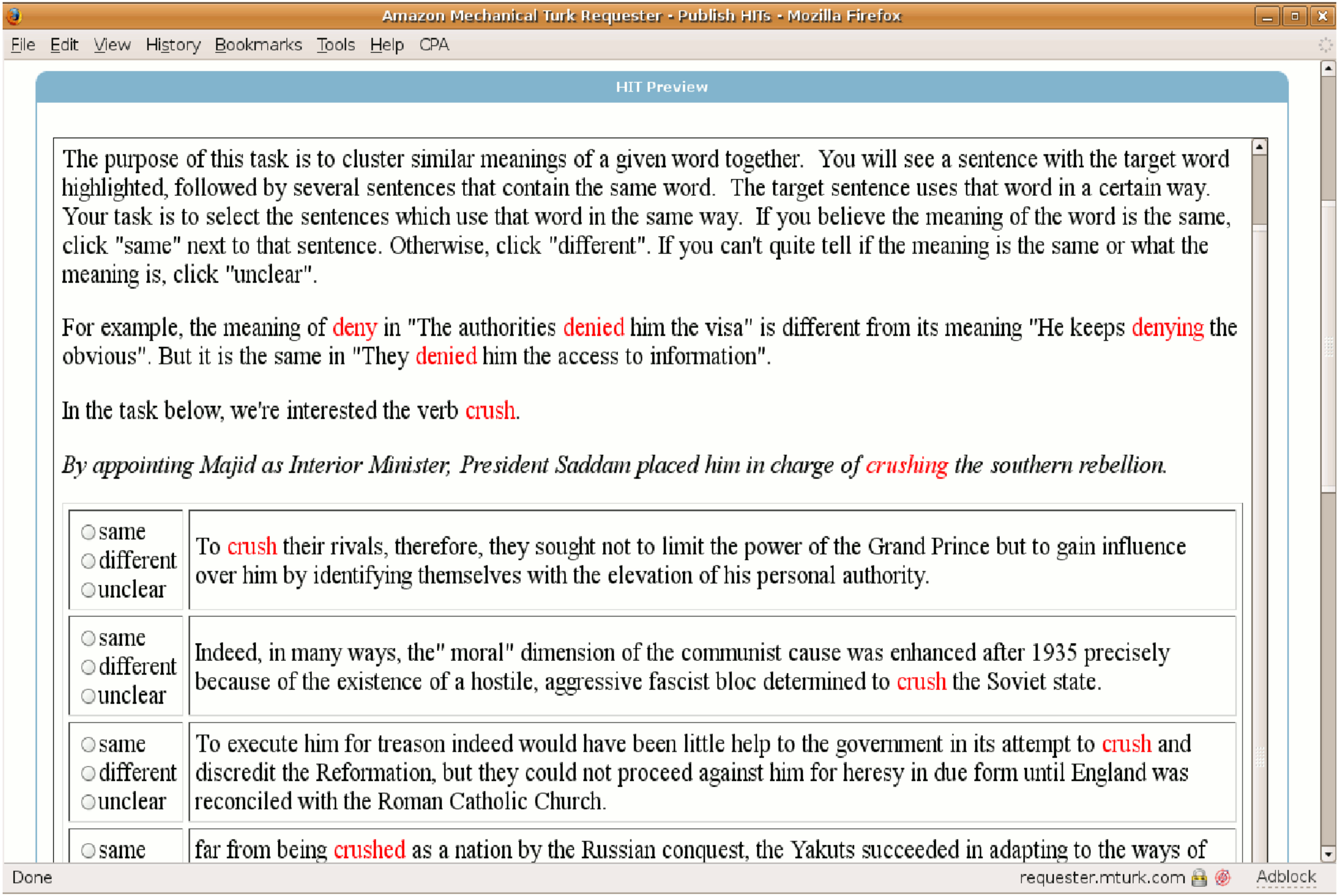
Annotation Task

- Sequence of annotation rounds, each round creating a cluster corresponding to a sense
- Turkers are given a set of sentences containing the target word, and one sentence that is randomly selected as the prototype sentence
- The task is to identify, for each sentence, whether the target word is used in the same way as in the prototype sentence

Proof of Concept Experiment

- Test verb: “crush”
- 5 different sense-defining patterns according to the CPA verb lexicon
- Medium difficulty both for sense inventory creation and annotation
- Test set: 350 sentences from the BNC classified by a professional lexicographer

Annotation Interface for the HIT


Amazon Mechanical Turk Requester - Publish HITs - Mozilla Firefox
File Edit View History Bookmarks Tools Help CPA
HIT Preview
Done
requester.mturk.com
Adblock

The purpose of this task is to cluster similar meanings of a given word together. You will see a sentence with the target word highlighted, followed by several sentences that contain the same word. The target sentence uses that word in a certain way. Your task is to select the sentences which use that word in the same way. If you believe the meaning of the word is the same, click "same" next to that sentence. Otherwise, click "different". If you can't quite tell if the meaning is the same or what the meaning is, click "unclear".

For example, the meaning of deny in "The authorities denied him the visa" is different from its meaning "He keeps denying the obvious". But it is the same in "They denied him the access to information".

In the task below, we're interested the verb crush.

By appointing Majid as Interior Minister, President Saddam placed him in charge of crushing the southern rebellion.

○ same ○ different ○ unclear	To crush their rivals, therefore, they sought not to limit the power of the Grand Prince but to gain influence over him by identifying themselves with the elevation of his personal authority.
○ same ○ different ○ unclear	Indeed, in many ways, the" moral" dimension of the communist cause was enhanced after 1935 precisely because of the existence of a hostile, aggressive fascist bloc determined to crush the Soviet state.
○ same ○ different ○ unclear	To execute him for treason indeed would have been little help to the government in its attempt to crush and discredit the Reformation, but they could not proceed against him for heresy in due form until England was reconciled with the Roman Catholic Church.
○ same	far from being crushed as a nation by the Russian conquest, the Yakuts succeeded in adapting to the ways of

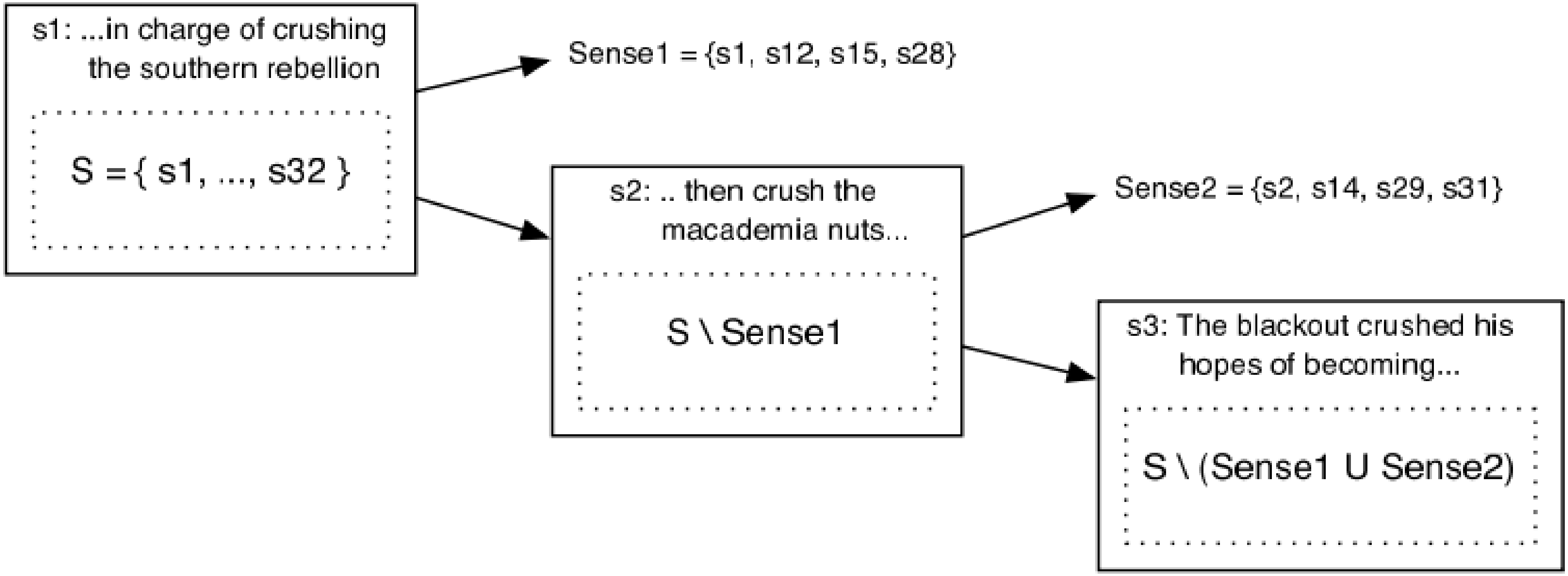
Annotation HIT Design

- 10 sentences per page
- Each page annotated by 5 different Turkers
- Self-declared native speakers of English

Annotation Task Rounds

- After the first round is complete, sentences judged as similar to the prototype by the majority vote are set apart into a separate cluster corresponding to a sense and excluded from further rounds
- The procedure repeated with the remaining set, i.e. a new prototype sentence selected at random, and the remaining examples presented to the annotators

Annotation Task Rounds


s1: ...in charge of crushing the southern rebellion
S = { s1, ..., s32 }
Sense1 = {s1, s12, s15, s28}
s2: .. then crush the macademia nuts...
S \ Sense1
Sense2 = {s2, s14, s29, s31}
s3: The blackout crushed his hopes of becoming...
S \ (Sense1 U Sense2)

Annotation Task Rounds

- The procedure is repeated until no examples remain unclassified, or all the remaning examples are classified as unclear by the majority vote
- Since some misclassifications are bound to occur, we stopped the iterations when the remaining set contained 7 examples, judged by an expert to be misclassifications

Annotation Procedure and Cost

- One annotator completed each 10-sentence page in approx. 1 min
- Annotators work in parallel
- Each round took approx. 30 min total to complete
- Annotators were paid $0.03 per page
- The total sum spent on this experiment did not exceed $10

Output for “crush”

- Three senses, with the corresponding clusters of sentences
- Prototype sentences for each cluster:
  - By appointing Majid as the Interior Minister, President Saddam placed him in charge of crushing the southern rebellion
  - The lighter woods such as balsa can be crushed with finger
  - This time the defeat of his hopes didn't crush him for more than a few days

Evaluation

- Against a gold standard of 350 instances created by a professional lexicographer for the CPA verb lexicon
- Evaluated using the standard methodology used in word sense induction (cf. SemEval-2007)
- Will refer to
  - Clusters from the gold standard are as *sense classes*
  - Clusters created by non-expert annotators as *clusters*

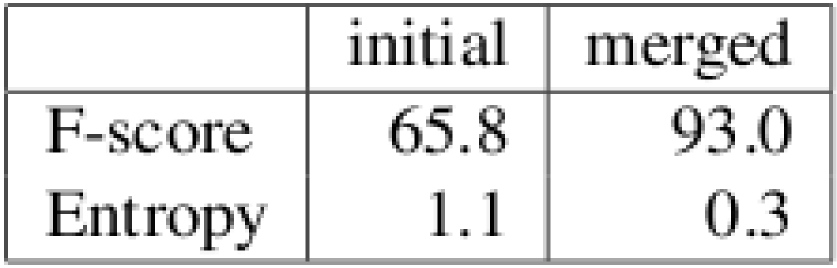
Evaluation Measures

- Set-matching *F-score* (Zhao et al, 2005; Agirre and Soroa, 2007)
  - Precision, recall, and their harmonic mean (F-measure) computed for each cluster/sense class pair
  - Each cluster paired with the class that maximizes it
  - F-score computed as a weighed average of F-scores obtained for each matched pair (weighted by the size of the cluster)
- *Entropy* of a clustering solution
  - Weighted average of the entropy of the distribution of senses within each cluster

$$\mathbf{Entropy}(C,S) = -\sum_i \frac{|c_i|}{n} \sum_j \frac{|c_i \cap s_j|}{|c_i|} \log \frac{|c_i \cap s_j|}{|c_i|}$$

where $c_i \in C$ is a cluster from the clustering solution C and $s_j \in S$ is a sense from sense assignment S

Results

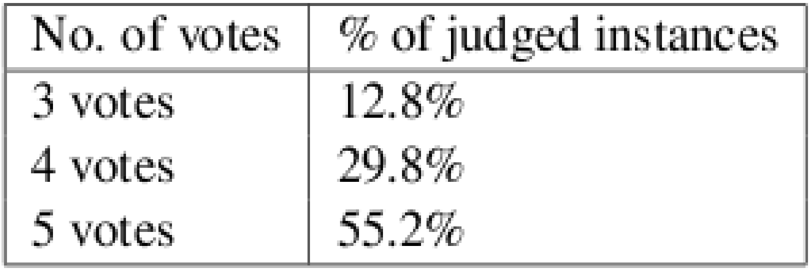


	initial	merged
F-score	65.8	93.0
Entropy	1.1	0.3

- Initial results figures compare 5 expert classes to 3 clusters
- CPA verb lexicon classes correspond to syntactic and semantic patterns, sometimes with more than one pattern per sense
- We examined the CPA patterns for crush, merged the pairs of classes corresponding to the same sense.
- Evaluation against the resulting merged classes is a near match!

Inter-Annotator Agreement

- Fleiss' kappa was 57.9
- Actual agreement 79.1 %
- Total number of instances judged 516
- Distribution of votes in majority voting:



No. of votes	% of judged instances
3 votes	12.8%
4 votes	29.8%
5 votes	55.2%

Issues and Discussion

- Annotators that perform poorly can be filtered out automatically, by throwing out those that tend to disagree with the majority judgement
- In our case, ITA was very high despite the fact that we performed no quality control!

Issues for constructing a full Sense-Annotated Lexicon

- Clarity of sense distinctions
  - Consistent sense inventories may be harder to establish for some words, esp. for polysemous words with convoluted constellations of related meanings (e.g. drive)
- Quality of prototype sentences
  - If sense of the target is unclear in the prototype sentence, quality of the cluster would fall drastically
  - This could be remedied by introducing an additional step, asking another set of Turkers to judge the clarity of the prototype sentences
- Optimal number of Turkers
  - Five annotators may not be the optimal figure
- Automating quality control and subsequent HIT construction

Conclusions and Future Work

- Empirically-founded sense inventory definition
- Simultaneously producing sense-annotated corpus
- Possible problems
  - Polysemous word with convoluted constellations of meaning, e.g. drive
- Evaluate against other resources
- Does not resolve the issue of task-specific sense definition
- But: a fast and cheap way to produce reliable, generic, empirically-founded sense inventory!

More Complex Annotation Tasks?

- CPA
  - [[Anything]] crush [[Physical Object = Hard | Stuff = Hard]]
  - [[Event]] crush [[Human | Emotion]]
- Argument Selection and Coercion / GLML (Semeval-2010)

  Sense 1

  - The general denied this statement (selection)
  - The general denied the attack (Event → Prop / coercion)

  Sense 2

  - The authorities denied the visa to the general

Thank you!