



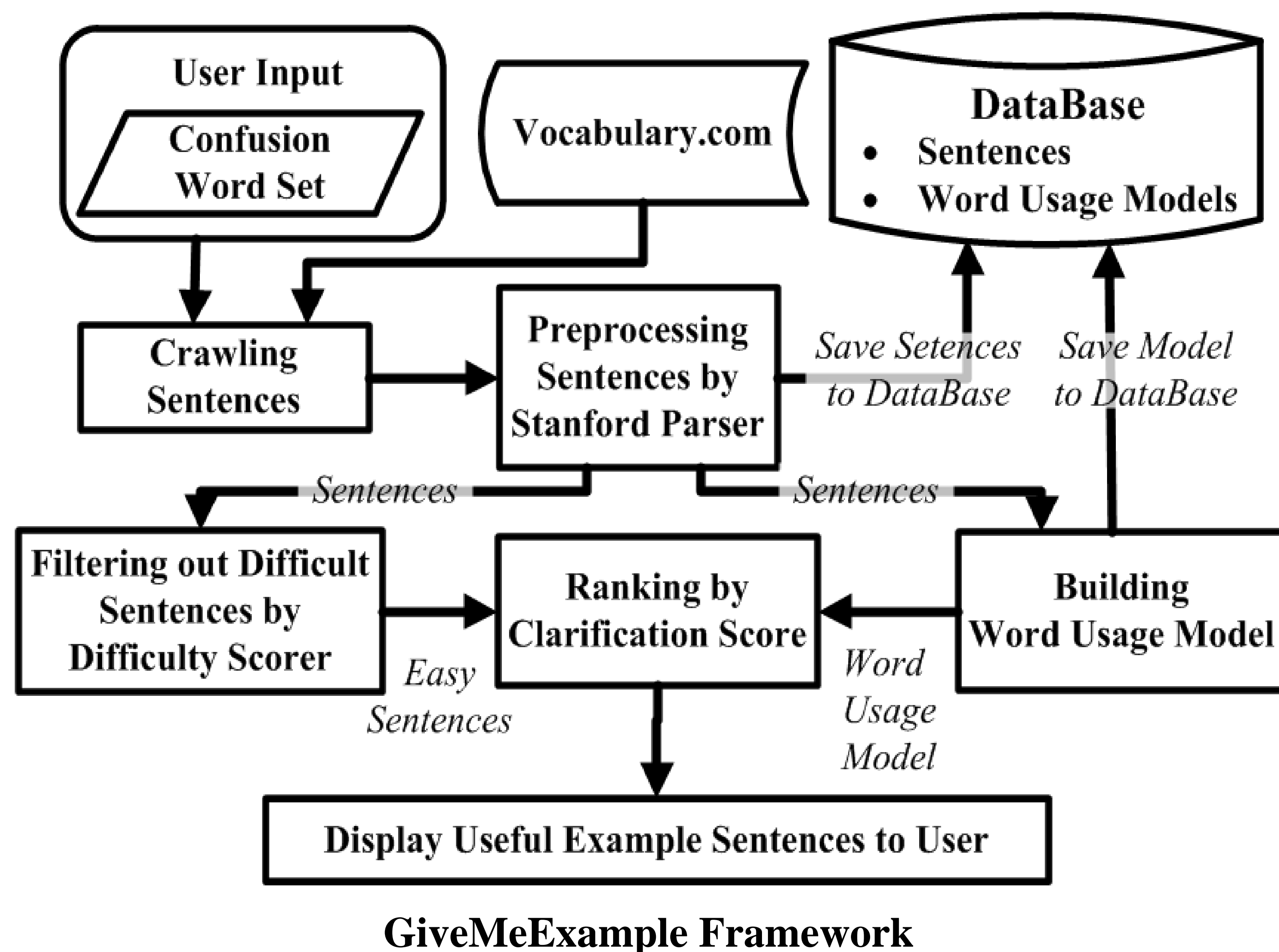
GiveMeExample: Learning Confusing Words by Example Sentences

Chieh-Yang Huang, Lun-Wei Ku

Institute of Information Science, Academia Sinica

Contact: appleternity@iis.sinica.edu.tw, lwku@iis.sinica.edu.tw

SYSTEM FRAMEWORK



EXPERIMENT

- Learners labeled dataset.

Refuse	Disagree
1 Police say Holmes emerged from a stolen car and refused to drop a gun before being shot by police.	1 In the report, he disagreed that the band's culture is sexualized.
2 Mr. Sentsov said he had been pressed to confess but refused .	2 "We disagree with ruling and are considering our options," a Morgan Stanley spokeswoman said.
3 Since then, service providers have refused in many cases to provide emails without warrants.	3 He disagreed with the premise that The Times's interest has been anything less than intense.
4 Both the insurer and tax board have refused to publicly release the audit and related records.	4 I don't subscribe to the pope's religion, and I disagree with many of his positions.
5 Putin has refused to contact Erdogan over the incident.	5 Justice Kagan disagreed , writing that "a 'tangible object' is an object that's tangible".

An Example Question We Provide To Learner

- A total of 6 learners of 2 high, 2 medium, 2 low language proficiency help us label the most useful example sentence pair.
- 10 verb confusion sets and 200 questions are proposed for evaluation.
- For each question, we use the best rank among the 6 gold pairs to calculate the Mean Reciprocal Rank (MRR).

$$MRR = \frac{1}{||Q||} \sum_{q \in Q} \frac{1}{\min(rank_q)}$$

DIFFICULTY SCORER

- The automatic difficulty scorer is built based on the work of Pisan et al. but with several modifications.
- In order to give a score to a sentence instead of a category, we apply linear regression instead of SVM.
- Some features only for the Swedish language in Pisan's work are removed.
- Training data is manually labeled by a native speaker, who considers the degree of difficulty of composite lexicons and the syntactic structure of the sentences.
- We set hard constrain for the difficulty score using a upper bound threshold.
- "I accept it." is simple but not a good example sentence because it only provides limited information.

WORD USAGE MODEL

- We build word usage model to estimate $P(s|w)$ for each word w with the observed sentence s .
- Contexture Feature
 - Given an observed sentence $s = w_1 \dots w_{i-k} \dots w_i \dots w_{i+k} \dots w_n$, where w_i is target word and k is window size.
 - Contexture Feature will be $\{e_{w_{i-k}} P_{w_{i-k}} \dots e_{w_{i-1}} P_{w_{i-1}} e_{w_{i+1}} P_{w_{i+1}} \dots e_{w_{i+k}} P_{w_{i+k}}\}$, where e_w is word embedding of w , P_w is Part-Of-Speech of w .
- We apply Gaussian Mixture Model (GMM) to learn data's distribution.
- For each word, we use 5000 sentences to train its word usage model. Besides, the number of Gaussian mixture model is set to 50.

LEARNING DIFFERENCE

- When searching for the useful example sentences of the target word w_i in word set W . The clarification ability relates to two factors.
- Fitness Score:** $P(s|w_i)$, whether w_i is appropriate for the sentence s .
- Relative Closeness:**
 $\prod_{w_j \in W - w_i} P(s|w_i) / P(s|w_j)$, multiplication of probability ratios. The idea comes from that s should fit the target word w_i but be inappropriate for the rest of words in W .
- We define the clarification scoring function as the multiplication of these two scores:

$$score(s|w_i) = P(s|w_i) * \prod_{w_j \in W - w_i} \frac{P(s|w_i)}{P(s|w_j)}$$

RESULT

	accept / agree	delay / postpone	refuse / disagree	excuse / forgive	invent / discover
GiveMeExample	0.547	0.549	0.560	0.508	0.376
Random	0.434	0.428	0.413	0.402	0.411
	manage / arrange	prevent / deter	realize / understand	destroy / spoil	occur / happen
GiveMeExample	0.456	0.420	0.511	0.447	0.450
Random	0.428	0.433	0.438	0.419	0.424

- Average MRR of GiveMeExample is 0.486.

- Average MRR of Random Baseline is 0.423.



<http://givemeexample.com/GiveMeExample>