Extraction of Example Sentences for Improved Reading and Understanding of Japanese Texts

Nahian Jahangir

2015

The Ambiguous Nature of Language

"Mary **engaged** Tom with her veritable knowledge of petunias and her sleight of hand tricks."

Definition of Engage:

- (verb) to occupy the attention or efforts (of people)
- (verb) to betroth
- (verb) enter into conflict with

Solution? **Example Sentences**

For example:

"The children **engaged** the teacher by asking several questions about the subject."

Advantages

- Simpler terms
- Retains context
- Uses basic grammar

Why Japanese?

"大学の食堂でハンバーガを食べてもいいですか？"

Is it okay if we eat hamburgers at the college cafeteria?

Difficulties of Japanese Language

- Three different writing systems: hiragana ひらがな, katakana カタカナ, and kanji 漢字.
- Heavily context-based language
- Level 4 Language- classified by DLI[1]

1: The defense language institute. http://new.dliflc.edu/. Accessed: 10/05/2015.

Applying it to Japanese

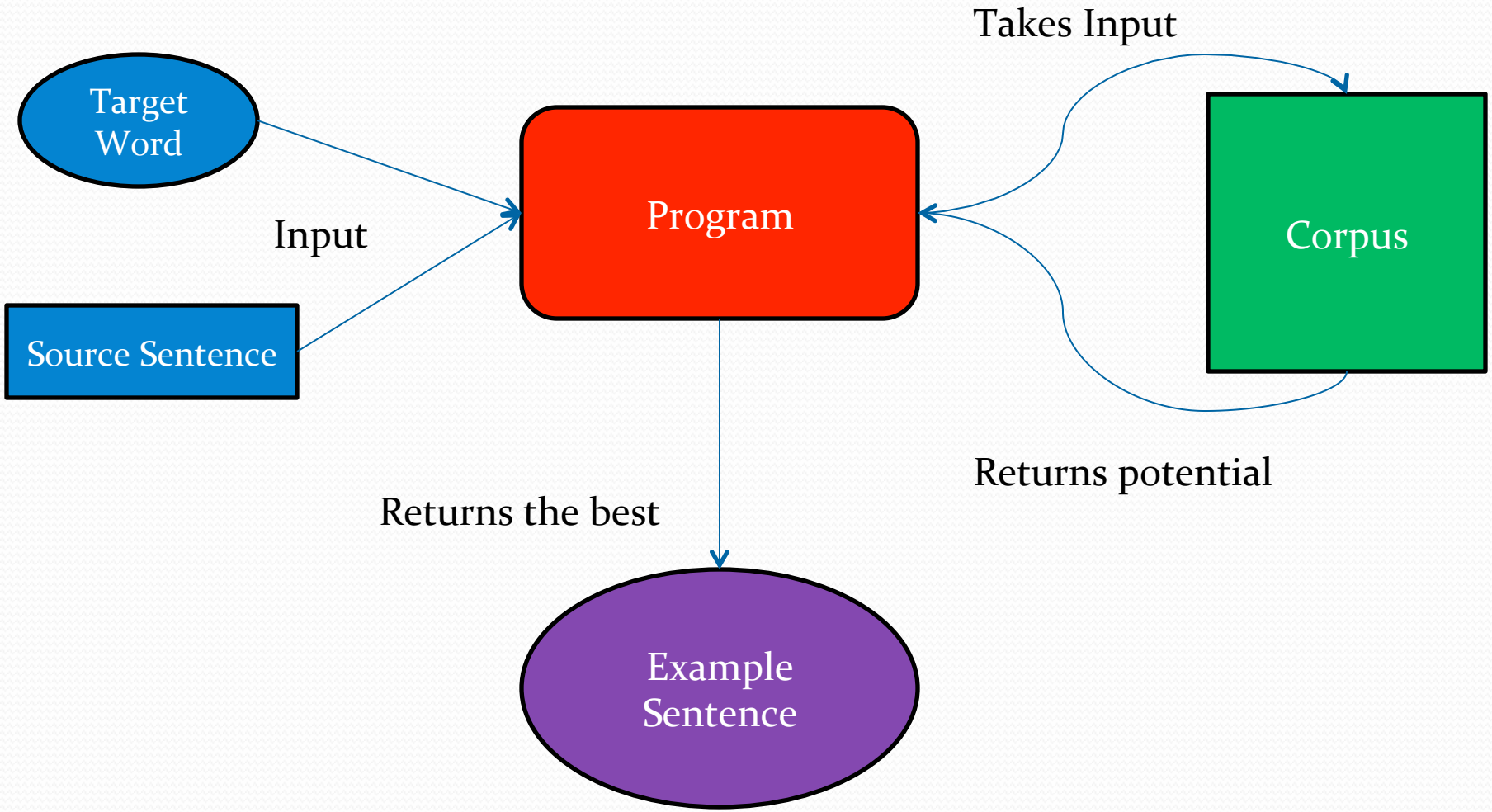
両親が年をとったら面倒を見るつもりです

- 面倒 [mendou] (Na- adjective, noun)*
  - Trouble; Difficulty; Care; Attention
- **Context Examples**
- 面倒を見る - to care for someone; to look after someone
- 面倒を掛ける - to put someone to trouble

あなたが買物に行っている間、子供の面倒を見ましょう

*Definitions courtesy of jisho.org, an online electronic dictionary server

How would it work?



The Tanaka Corpus[2]

Characteristics

- Multi-lingual parallel corpus of English and Japanese
- Sentences were every day use sentences
- Edited and corrected for mistakes

Further alterations

- Removed English sentences and duplicate, formatted sentences
- From 420,000 sentences → **149,298 sentences**

2: Electronic dictionary research and development group. http://www.edrdg.org/wiki/index.php. Accessed: 21/11/2014.

The LESK Algorithm

Overview[3]

- Introduced by Michael E. Lesk in 1986
- Derives from word sense disambiguation

Problems

- Need exact definitions
- Limited to dictionary glosses

Solution-Simplified Lesk Algorithm

3: Agirre, Eneko, and Philip Edmonds. "Word Sense Disambiguation: Algorithms and Applications."

Simplified LESK Algorithm

function SIMPLIFIED LESK(*word,sentence*) **returns** *best sense of word*

best-sense ← *most frequent sense for word*

max-overlap ← *0*

context ← *set of words in sentence*

for each *sense* **in** *senses of word* **do**

signature ← *set of words in the gloss and examples of sense*

overlap ← COMPUTEOVERLAP (*signature,context*)

if *overlap* > *max-overlap* **then**

max-overlap ← *overlap*

best-sense ← *sense*

end return (*best-sense*)

Baseline

function OVERLAP(word, sentence) **returns** *best example sentence*

best score ← 0

example sentence ← ””

source vector ← SETCREATION(*sentence*)

for *other sentence* **in** *corpus* **do**

other vector ← SETCREATION(*other sentence*)

if *word* **in** *other vector* **then**

score ← COMPARE_OVERLAP(*vector*, *other vector*)

if *score* > *best score* **then**

best score ← *score*

example sentence ← *other sentence*

end return *example sentence*

Test Sentences

- 両親
  - 両親が年をとったら面倒を見るつもりです
  - Translation: In the case my parents get older with age, I will look after them.
- 頼む
  - 夕方になると忙しくなるから、頼むよ
  - Translation: In the evening it will get busy, so I am counting on you.
- 安い
  - バスと電車とどっちのほうが安いですか
  - Translation:　Which is cheaper, (going by) bus or (by) train?

11 sentence: もう少し安い部屋がありますか。 score: 0.333333333333

12 sentence: それは安いですね。 score: 0.333333333333

13 sentence: ９時以降に電話した方が安いですか。 score: 0.307692307692

14 sentence: 二つのうちではこちらの方が安い。 score: 0.3

15 sentence: 値段が安いのはうれしい驚きだった。 score: 0.3

16 sentence: 国内便の安い航空券はありますか。 score: 0.3

17 sentence: そのネックレスが１００ドルとは安い。 score: 0.3

18 sentence: この季節は卵が安い。 score: 0.285714285714

19 sentence: 結局はツアーに入っちゃうのが安いよね。 score: 0.272727272727

20 sentence: 神戸は比較的物価が安い。 score: 0.25

21 sentence: もっと安い部屋はありますか。 score: 0.25

22 sentence: もっと安いものはありますか。 score: 0.25

Baseline Results

- 両親（**105**）
  - 9,13 23, 27, 44, 58, 60, 71, 73,
- 頼む（**27**）
  - 11, 13 ,23
- 安い (**80**)
  - 15, 17, 29, 38, 43, 63, 64, 67, 68, 74, 79

Evaluations

- Longer sentences hold unfair advantage
  - Normalization solves for this

Improvements/Approaches

- Remove stop words (particles)
- Collocation of Sentences (Method #1)

Method #1 Results

- 両親（**105**）
  - 3, 10, 25, 28, 34, 38, 45, 52, 93
- 頼む（**27**）
  - 14, 16, 27
- 安い (**80**)
  - 7, 16, 21, 23, 37, 40, 44, 47, 56, 64, 68

Evaluations

- Overall scores were generally higher
  - Collocations based on common phrases found throughout corpus → higher scores given to them

Improvements/Approaches

- Weighting the words (Method #2)

Method #2

- 両親（**105**）
  - 8, 10, 11, 35, 43, 65, 81, 84, 85, 86,
- 頼む（**27**）
  - 17, 22, 25
- 安い (**80**)
  - 17, 27, 29, 30, 34, 40, 54, 59, 65, 66, 69

Evaluations/Discussion

- Longer sentences had more opportunities to score higher
  - Try normalizing

Future Work

- Normalization and stop character removal
- Incorporate Kanji Proficiency
- Continuing Method 1
- Including more corpora