

要求仕様書における重要用語語義曖昧性の位置付けと解釈候補の可視化

松岡 仁 ルパージュ イヴ
早稲田大学大学院 情報生産システム研究科

jinmatsuoka@akane.waseda.jp yves.lepage@aoni.waseda.jp

1 はじめに

多くの産業では、QCD (Quality, Cost and Delivery) という3つの要因で製品の生産性を測る指針として用いられている。品質 (Quality) が良い製品を作ることができたか、どれだけ低い経費 (Cost) で製品を作り上げたか、納期 (Delivery) に間に合うように製品を作ることができたか、という3つの指針となる。ソフトウェア開発でも同様な指針を用いられている。

ソフトウェア開発での超上流の要求工程で生成される要求仕様書がどれだけの曖昧性を除去できるかで、プロジェクトの成功や失敗に関わる重大な問題になる。そのため、IEEE std 830-1998¹ という定義書があり、詳細な記述方法が定義されている。ソフトウェア要求仕様書は自然言語で記述される以上、曖昧性を含む可能性がある。従って、早い段階で曖昧性を検出し、対応しなければならない。

自然言語処理では語義曖昧性解消は情報検索、情報抽出と機械翻訳 [1] での体系の一部として位置付けられている。関連している研究として [4] があるが、本論文ではその違いとして階層的な意味的ネットワーク上で類似度計算を行う手法を用いる。

本論文では、第2章にどのような言語資源と技法を用いたのかを説明する。そして、第3章で実験データと実験手順を説明し、結果を示す。最後に結論とする。

2 言語資源と技法

2.1 WordNet

WordNet は階層的な語彙ネットワークである。本研究では日本語仕様書について対象とするので、プリンストン大学の英語 WordNet [6] ではなく、情報通信研究機構 (NICT)² で開発された日本語 WordNet [3] 用

¹<http://standards.ieee.org/findstds/standard/830-1998.html>

²<http://nlpwww.nict.go.jp/wn-ja/index.en.html>

いた。

WordNet 上のノードは同義語集合 (Synset) である。曖昧のある単語は WordNet 上で2つ以上のノードに依存する。例えば、日本語の「いか」には複数の意味がある。「空に飛ばすおもちゃ」と「食べもの」の2つの意味である³。単語レベルでは、どちらかの意味で使用されているかは決定することはできないが、文脈を用いて決定することができる。以下にその例を示す。

Sentence	Meaning	
	‘おもちゃ’	‘食べ物’
私は公園で/いか/をあげる	yes	no
私は油で/いか/をあげる	no	yes

本研究では、WordNet 上で「いか」の意味を決定するために、「いか」についての上位語/下位語のリンクのみを考慮する。

2.2 WordNet での意味的類似度

WordNet の toolbox には概念間 (Synset) 同士の意味的類似度を測る尺度が提供されている。4つの意味的類似度を測る方法があるが、本論文では Wu と Palmer が考案した wup_similarity [7] を用いる。以下にその式を示す。

$$\text{wup_similarity}(s_1, s_2) = \frac{2 \times \text{depth}(\text{lcs}(s_1, s_2))}{\text{depth}(s_1) + \text{depth}(s_2)} \quad (1)$$

s は Synset であり、depth は最上位概念までの距離である。LCS (Lowest Common Synset) という最小共通概念である。考え方は Dice 係数と同じであるが、Dice 係数は集合論に基づいているが wup_similarity は概念階層に基づいている。この式を図1で示す。

上記の式を用いることで用語の意味決定を行うための式を下記に示す。

$$\text{word_sim}(w_1, w_2) = \max_{(s_i, s_j)} \text{wup_similarity}(s_i, s_j) \quad (2)$$

³関西圏では、/たこ/のことを/いか/とも表現される

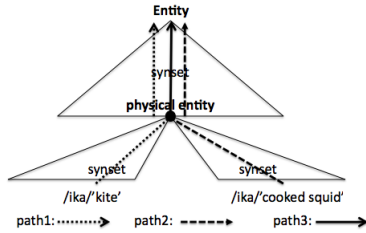


図 1: 意味的類似度の計算方法: /いか/ 'おもちゃ' と /いか/ '食べ物' の 2 つの Synsets の最小共通概念は *physical entity* である. 類似度の値は $2 \times \text{length}(\text{path3}) / (\text{length}(\text{path1}) + \text{length}(\text{path2}))$ である.

w_1 と w_2 は単語であり、それらの単語の語義 (Word Sense) の Synset を WordNet より全て抽出する. w_1 と w_2 を選択する際の条件は下記になる.

$$w_1 \in s_i \wedge w_2 \in s_j \wedge \text{pos}(w_1) = \text{pos}(w_2) \quad (3)$$

POS は品詞 (Part Of Speech) である. そして、これらの Synset との文にある単語との意味的類似度を *wup_similarity* で測り、最適値を求める. その式を下記に示す.

$$\text{sem_val}(s, S) = \max_{s_j/w_i \in s_j \wedge w_i \in S \setminus \{w\}} \text{word_sim}(s, w_i) \quad (4)$$

上記の式を実行するとそれぞれの意味に対し、その単語と概念的の類似度を計算できる. 上記の式を用いて、例として、/いか/について「私は/いか/を公園であげる」での計算結果を下記に示す (この例では事前に分かち書きを施している).

	私	は	公園	で	/いか/	を	あげる
'おもちゃ'	0.00	-	0.56	-	-	-	-
'食べ物'	0.00	-	0.30	-	-	-	-

上記の例で最適値をとれば、その文での曖昧の用語がどのような意味で使われているかを決定するための式を下記に示す.

$$s = \arg \max_{s_i/w \in s_i} \text{sem_val}(s_i, S) \quad (5)$$

これらの式を用いて、「たこ」を含む他の文にも計算を行い、下記の表 1 に示す.

2.3 語義の意味による文クラスタリング

第 2.2 章で紹介した手法を用いて、「いか」を含む複数の文を分類する. ここでは、クラスターは WordNet から抽出した「いか」の語義 (意味) となる.

表 1: 「いか」を含む複数文での計算結果

Sentences	Computed meaning
私は公園で/いか/をあげる.	{'おもちゃ'}
私は高台で/いか/をあげる.	{'おもちゃ', '食べ物'}
私は油で/いか/をあげる.	{'食べ物'}
夕食は/いか/だ.	{'食べ物'}
私は水で/いか/をゆでる.	{'おもちゃ', '食べ物'}
私は/いか/を食べる.	{'おもちゃ', '食べ物'}

表 2: 表 1 をそれぞれの意味クラスターで考えた分類クラスター (Cluster) 文 (Sentences)

クラスター (Cluster)	文 (Sentences)
{'おもちゃ'}	私/は/広場/で/いか/を/あげる/.
{'食べ物'}	私/は/油/で/いか/を/あげる/. 夕食/は/いか/だ/.
{'おもちゃ', '食べ物'}	私/は/高台/で/いか/を/あげる/. 水/で/いか/を/ゆでる/. いか/を/食べる/.

上記の表で 2 つの意味に割り当てられたクラスターの中に 3 つの例文の中があり、その一例を下記に示す.

	私	は	高台	で	/いか/	を	あげる
'おもちゃ'	0.00	-	0.13	-	-	-	-
'食べ物'	0.00	-	0.13	-	-	-	-

上記の表で同じ最適値になり、意味を決定することができない. 可能な限り 1 つの意味に決定するための手法を第 2.4 章で説明する.

2.4 文と文との意味的類似度

第 2.3 章で挙げた問題を解決するために、文と文との意味的類似度 [2] を計算する. 下記に文と文との意味的類似度を計算するための式を示す.

$$\text{sent_sim}(S_1, S_2) =$$

$$\frac{1}{2} \times (\text{sem_div}(S_1, S_2) + \text{sem_div}(S_2, S_1))$$

S_1 と S_2 は文 (sentence) であり、 $\text{sent_div}(S_1, S_2)$ は非対称であるために算術平均を求める. 次に sent_div の式を下記に示す.

$$\text{sem_div}(S_1, S_2) =$$

$$\frac{\sum_{w_i \in S_1} \left(\max_{w_j \in S_2} \text{word_sim}(w_i, w_j) \times \text{idf}(w_i) \right)}{\sum_{w_i \in S_1} \text{idf}(w_i)}$$

S_1 と S_2 は文の中に出現する単語を w とする. この式の値域として $0 \leq \text{sent_div}(S_1, S_2) \leq 1$ である. この式では、 S_1 と S_2 との重心を求めている. 考え方として min-max centering 手法である. そして、重心を求める

ための重みとして idf (inverse document frequency) の単語と文書の出現数の自己情報量を用いている。

$$\text{idf}(w) = -\log \frac{|\{S \in D/w \in S\}|}{|D|}$$

D は文書 (document) であり、 $|D|$ は文書の数である。これらの計算式を用いて、単語と単語のアライメント手法のように文と文との意味的類似度を計算する。

2.5 分割クラスタリング

第2.4章の文と文との意味的類似度を計算し、文での分割クラスタリングを行う。分類した結果を下記の表に示す。

Cluster	Sentences
{‘おもちゃ’}	私/は/公園/で/いか/を/あげる/。 私は高台/で/いか/を/あげる/。 私は水/で/いか/を/ゆ/でる/。
{‘食べ物’}	私/は/油/で/いか/を/あげる/。 夕食は/いか/だ。
{‘おもちゃ’, ‘食べ物’}	私は/いか/を/食べる/。

ここでは、文と文との意味的類似度に閾値を設け、0.7 以上であれば、その文と文は等価に扱うことができる。

2.6 用語抽出

要求仕様書から曖昧用語をどのように抽出するかを説明する。重要な用語は専門用語も含むという観点から、C-value 手法 [5] を用いて、専門用語を抽出する。C-value 手法は言語的、統計的手法と位置付けることができる。まず、計算式を下記に示す。

$$\text{C-value}(a) = \begin{cases} \log_2 |a| \times \left(f(a) - \frac{1}{|T(a)|} \sum_{b \in T(a)} f(b) \right) \\ \log_2 |a| \times f(a) \text{ if } T(a) = \emptyset \end{cases}$$

ここでは、 a は専門用語となり、 $f(a)$ でその専門用語が文章中に何回出現しているかであり、 $|T(a)|$ は a の専門用語が他の専門用語の部分として現れる数である。単語の出現頻度と長さで重要な専門用語を抽出している。

3 実験と結果

3.1 実験データ

本研究で参照している文書は「照明装置制御仕様書」である。ページ数は 20 ページで、全文は 1177 文であ

表 3: C-value 手法を用いて専門用語を抽出した結果 (右端は WordNet に含まれているか、含まれていないか)

Extracted term	C-value	WordNet?
照明装置	86.1	
ネットワーク制御装置	69.7	×
ネットワーク制御装置無線モジュール	30.2	×
装置システムパラメーター報知メッセージ	27.3	×
装置システムパラメーター	26.9	×
装置無線モジュール	24.8	×
装置登録要求メッセージ	18.7	×
照明装置システムパラメーター報知メッセージ	18.1	×
照明装置システムパラメーター	16.7	×
メッセージ全体	14.0	×

る。特徴としては、数式を多く使い、片仮名表記が多かった点である。

表3に要求仕様書に対して C-value 手法を実施し、得られた結果を示す。

これらの専門用語の多くは WordNet に存在しないため、それらを区切り用語として実験では扱う。専門用語をからなる用語も重要である。

3.2 実験手順

下記に行った 6 つの実験手順を示す。

1. 文書を文単位と段落単位で抽出する
2. 得られた文を分かち書きと品詞付けを行う (形態素解析器 JUMAN⁴の利用)
3. C-value 手法で専門用語とそれからなる単語を抽出する
4. それぞれの専門用語と用語をを含む文で用語の意味決定クラスタリングを行う
5. 1 つの意味に決定できない専門用語や用語に対して、分割クラスタリングを行う
6. 意味決定で得られた用語に関して、2 つ以上の意味があるかないかを検証する

これら 6 つの実験手順に従い、得られた結果について第3.1章で検証する。

3.3 実験結果

第3.2章で説明した実験手順に従って、実験を行い、得られた結果について、表 4に示す。表 4から実験結

⁴<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN>

表 4: それぞれの用語に関して意味決定プロセスの結果

Term	WordNetでの語義数	文数	クラスター数	クラスター毎語義数	IEEE 830 compliant
照明装置	1	48	1	1.0	ok
ネットワーク	4	39	2	2.0	-
制御	10	46	5	1.0	-
装置	6	74	2	2.7	-
無線	4	40	1	1.0	ok
モジュール	3	19	1	1.0	-
システム	8	23	2	1.5	-
パラメーター	4	16	3	1.3	-
登録	4	10	1	1.0	ok
要求	20	22	4	1.3	-
メッセージ	4	42	2	2.0	-
システム	8	27	4	1.3	-
報知	2	14	1	1.0	ok
照明	5	50	4	1.3	-
全体	5	29	1	1.0	ok
データ	3	12	1	1.0	ok
設定	9	24	4	2.5	-

果について検証する。例えば、「無線」では要求仕様書内では 40 文あるが、全ての文で 1 つのクラスターで 1 つの意味に割り当てることができる。従って、IEEE std 830-1998 compliant では ok となる。

しかし、「ネットワーク」では要求仕様書内では 39 文あり、提案した手法では 1 つのクラスターで 2 つの意味に割り当てられる結果となる。従って、IEEE std 830-1998 compliant では ok とはならない。

実験結果により、自動抽出した 17 の位置付けた曖昧用語の中、33% の場合一意の意味決定でき、残る 66% の場合に対して、複数の意味で用いられていると検証することができる。

4 おわりに

本研究では、要求仕様書における重要用語語義曖昧性を位置付けるための組合せ手法を提案し、実装を行った。

本研究で利用した言語資源、技法と提案した手法を以下にまとめる。

- 言語資源: WordNet
- 既存技法
 - 意味的類似度: wup_similarity, 文と文との意味的類似度
 - 専門用語抽出: C-value

- 本論文で提案した手法

- 語義の意味クラスタリングと分割クラスタリングの併用

予備実験で、実際の開発で使用された「照明装置制御仕様書」を用いて、同音用語による重要用語語義曖昧性の位置付けと解釈候補の可視化アプリケーションを構築し、検証を行った。

最後に、検証するための指針として IEEE std 830-1998 の定義書に従い、20 ページの要求仕様書の予備実験結果、自動抽出した 17 の位置付けた曖昧用語の中、33% の場合一意の意味決定でき、残る 66% の場合に対して、読み手の判断支援のために解釈候補の可視化アプリケーションを実装した。

参考文献

- [1] Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. *In The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, 2007.
- [2] Courtney Corley and Rada Mihalcea. Measuring the semantic similarity of texts. *EMSEE '05 Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, 2009.
- [3] Kiyotaka Uchimoto Takayuki Kuribayashi Francis Bond, Hitoshi Isahara and Kyoko Kanzaki. Extending the Japanese wordnet. *In 15th Annual Meeting of The Association for Natural Language Processing*, 2009.
- [4] Anne De Roeck Bashar Nuseibeh Hui Yang, Alistair Willis. Automatic detection of nocuous coordination ambiguities in natural language requirements. *in Proceedings of the IEEE/ACM international conference on Automated software engineering*, pages 53–62, 2010.
- [5] Hideki Mima Katerina Frantzi, Sophia Ananadou. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3:115–130, 2000.
- [6] George A. Miller. Wordnet: A lexical database for English. *Communications of the Association for Computing Machinery*, 38(11):39–41, 1995.
- [7] Zhibiao Wu and Martha Palmer. Measuring the semantic similarity of texts. *EMSEE '05 Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, 2009.