複合語構成規則を用いた意味解析

顔加軍，姜沛林，黒岩真吾，任福継
{yanjj, jiang, kuroiwa, ren}@is.tokushima-u.ac.jp
徳島大学　工学研究科

概要
本稿では、HowNet の information structure を基にした中国語の複合語構成規則により、複合名詞を自動合成した上、意味解析モデルを構築し、複合名詞の意味クラスを推定する。規則には統語情報、意味情報、依存情報、及び用例の単語がつけてあるのは本文の特徴である。意味クラスの特定に用いるのは類似度である。シソーラス HowNet に於けるノード間の距離で単語間の類似度を計算し、多義複合語に関する意味的な曖昧性を解消する。
キーワード：複合語　information structure　意味解析

１、はじめに

中国語の文を形態素解析した後、複合語の境界が不明である。複合語の境界を明確にした後、複合語内の依存構造、タグ付け情報、及び意味情報を推定するのは難しい。意味解析における曖昧性は二つに分けられる。一つは、多義語に関する意味的曖昧性である。もう一つは、依存関係にある２語の間の意味関係に関する意味的曖昧性である。本稿は中国語の言語処理に於ける複合語内のタグ情報、意味情報及び依存構造、の推定に目指すものであり、提案手法は information structure を基にした複合語構成規則により、意味解析モデルを構築し、複合語の意味クラスを推定する手法を提案する。本手法によって、複合語に関する曖昧性をある程度解消できる。

２、information structure を基にした複合語構成規則

２．１複合語構造の特徴

中国語の特徴の一つとしては、単音節の語が主で、語形変化を行わず孤立語的である。構成要素は以下の三つがある。

語幹（stem）：語の構成上、基幹的な役目をする自立語。

接頭辞（prefix）：もとの単語に意義上の修飾を加える付属語。

接尾辞（suffix）：意義を加えるとともに新たに別の品詞の資格を与える付属語。

これらの構成要素により複数の漢字からなる単語の結合形態としては派生語と複合語がある。

派生語：ある語幹を基として、接辞が付くことによって、別の一語となったもの。例：

prefix+stem　非/金属　超/音速　無/条件　過/飽和
stem+suffix　芸術/家　研究/員　可能/性　現代/化
prefix+stem+suffix　非/党/員　超/薄/型

複合語：本来それぞれ独立の構成要素が、二つ以上結合して、新たに単純な一語としての意味・機能をもつようになったもの。例：
stem+stem　結構/語法　交通/規則　抗震/救災
一つの漢字でも単語になりうるため、正規表現で書くと、

派生語：　$= prefix^{i} + \textbf{単語}^{+} + suffix^{j} (i + j >= 1)$
複合語：　$= \textbf{単語}\ + \textbf{単語}^{+}$
単語：　$= stem \mid \textbf{派生語} \mid \textbf{複合語}$

と表示できる。このような特徴の構造は造語性が高いため、自然言語処理において、膨大な未知語を生じさせる。情報が溢れる現代では、新しい単語をどんどん辞書に登録しても、全部登録するのは不可能である。またこの特徴により、複合語の意味クラスを推定するのも更に困難になる。我々は意味解析に困難をもたらす複合語の特徴を利用して、問題の解決に取り込む。

２．２Information structure

まず、我々は中国語のシソーラス HowNet から複合語の特徴と似合う information structure を基にして複合語構成規則を用意した。Information structure は以下のように、４つの部分からなっている。

SYN_S：A <-- {V <-- N}
SEM_S：(属性值) <-- {(事件,行动) <-- [施事] (人/拟人)}
Query1: 谁？
Answer1: A + V + N
Query2: 他（她）是做（干）什么的？
Answer2: A + V “的人”
EX：职业-经理-人,首席-观察-员,首席-执行-官,首席-合伙-人,

①統語構造部分 Syntactic structure （58）
SYN_S：A <— {V <— N}
②意味構造部分 semantic structure（271）
SEM_S：(属性値) <— {(事件,行动) <— [施事] (人/拟人)}
③統語分布部分 Syntactic distribution （49）
Query1: 谁？ Answer1: A + V + N
Query2: 他（她）是做（干）什么的？ Answer2: A + V “的人”
④Example 部分（11000）
EX 职业-经理-人,首席-观察-员,首席-执行-官,首席-合伙-人

２．３Information structure の拡張

既存の information structure に対して、人手で拡張を行う。
拡張１ syntactic structure に基づき、規則によって合成後のタグ情報を決め、文脈自由文法に拡張する。
拡張２ information structure にある依存構造を表す矢印によって複合語内の主辞要素（head）と補語要素（complement）の関係を明確化にする。

矢印が出るほうは主辞要素であり、H で表す。矢印の受けるほうは補語要素であり、C で表す。二つの要素が合成した後、傾向としては、複合語の意味素性はおおよそ主辞要素の意味素性と一致する。例えば、「首席-观察-员」の中に、既に複合語「观察-员」が存在する、「观察员」の意味情報は「员」の意味情報と一致する。解析結果の記述は以下のように、

```
(首席-观察-员SYN_S:NP=A <-- {V <-- N}  SEM_S:(属性值) <-- {(事件,行动) <-- [施事] (人/拟人)}  HC:
  (首席   SYN_S:A          SEM_S:(属性值)                                  HC:C)
  (观察-员SYN_S:NP=V<-N  SEM_S: (事件,行动) <-- [施事] (人/拟人) HC:H
    (观察SYN_S: V  SEM_S:(事件,行动)     HC:C)
    (员SYN_S: N    SEM_S: [施事] (人/拟人) HC:H)))
```

＝は導出、（ ）は意味素、｛ ｝は再帰的関係を表し、[ ] は event role（格フレームに相当）で、依存関係にある単語間の意味関係を表す。

特徴としては、我々の拡張後の複合語構成規則によって生成した解析結果には統語情報（SYN_S）意味情報（SEM_S）と依存関係（HC）が共に存在する。現段階の規模では、Information structure には 271 の規則がある。その内、複合名詞規則は 177（一般名詞 NP は 124、時間名詞 TNP は 23、数量名詞 CLASNP は 16、地名 PNP は 14、人名 NNP は 13、）複合動詞 VP は 72、複合形容詞 AP は 20、句 S は２という割合である。本稿はとりあえず複合名詞だけを取り扱う。

３、意味解析モデル

本節では、提案するモデルについてその概要を述べる。まず、概観を図１に示す。

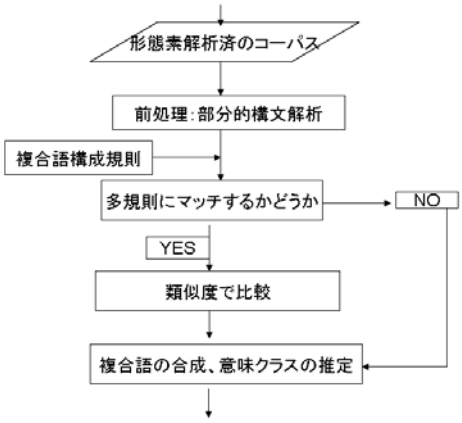

図１：意味解析モデル

前提としては、形態素解析済みのタグ付きデータを入力する。次に、複合語構成規則によって部分的構文解析し、隣り合う単語のあらゆる可能な合成を一つの組にし、準備された複合名詞規則とマッチする。いくつかの意味構造にマッチした場合、特定に用いるのは類似度である。このとき、規則にある用例とシソーラス HowNet に於けるノード間の距離で単語間の類似度を計算する。計算方法は LIU（2002）を用いる。
例えば：「图书/n 馆/n 」が最初にルール「NP＝N <— N」にマッチした。NP＝N <— N は相当に多い意味構造と対応している。例えば、同じ統語構造の「哲学家」「人事局」の意味構造は異なる。

SYN_S：NP＝N <— N SEM_S：(知识/信息）[内容]<—[施事]（人/组织/部件,%组织） EX:哲学-家

SYN_S :NP＝N <— N SEM_S：(物质/事情/事务)[受事]<—[施事]（组织/场所） EX:人事-局

次に、「图书」を「哲学」「人事」と、「馆」を「家」「局」と、或いは、直接「图书馆」と「哲学家」「人事局」との意味の類似性を計算する。意味素による単語定義辞書を検索すると、

图 书 馆 N InstitutePlace| 场 所 ,@read|读,@borrow|借入,#readings|读物

哲学家 N human|人,#knowledge|知识

人 事 局 N part| 部 件 ,%institution| 机构,*manage|管理,#employee|员

類似度でもっとも高い用例の統語構造と意味構造を抽出して単語に付与する。類似度計算のために、単語定義辞書の形式を書き直すと、図２のようになる。

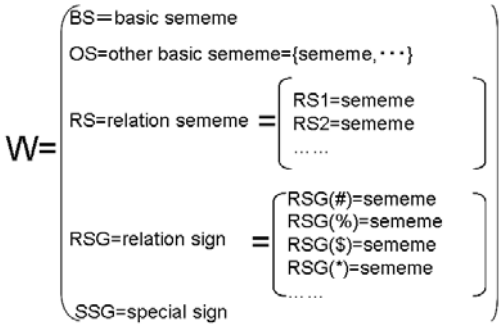

図２　単語定義辞書の形式

一つの単語の定義は５つの部分の意味素からなり、それぞれ基本意味素(BS)、その他の基本意味素(OS)、関係意味素（RS)、関係符号意味素（RSG)、特殊符号意味素(SSG)である。単語Ｗ１とＷ２との類似度は、5 つの部分の意味素の類似度$Sim_j(S_1,S_2)$の和からなる。式（１）に示す。

$$Sim(W_1,W_2)=\sum_{i=1}^{5}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2)$$
$$(\beta_1+\beta_2+\beta_3+\beta_4+\beta_5=1,\quad \beta_1\geq\beta_2\geq\beta_3\geq\beta_4\geq\beta_5) \qquad (1)$$

式（１）のように、単語の類似度は意味素の類似度に転換する。基本意味素(BS)の類似度$Sim_1(S_1,S_2)$は意味素シソーラスにおける距離dによって決められる。式（２）に示す。αは類似度が 0.5 になったときの距離である。

$$Sim_1(S_1,S_2)=\frac{\alpha}{d+\alpha} \qquad (2)$$

その他の基本意味素（OS）の類似度$Sim_2(S_1,S_2)$は二つの集合の類似度で表す。集合OS1とOS2に対し、まず、OS１から一番目の元素を取り、OS2 の全ての元素との類似度を計算し、その中から値がもっとも高い元素を抽出し、一対一の関係を確立する。一対一関係が確立できたものは集合から消す。対応する元素がない場合は、null と対応させ、類似度を値が小さい定数δにする。元素が全部消されるまで続く。一対一の対応関係が終了後、集合間の類似度は各元素対の類似度の算術平均からなる。式（３）に示す。そのうちｎとｍはそれぞれ二集合の元素数、ｘは共通な元素の数である。

$$Sim_2(S_1,S_2)=\frac{Sim(\rho_{n1},\rho_{m1})+Sim(\rho_{n2},\rho_{m2})+\ldots}{n+m-x} \qquad (3)$$

関係意味素（RS)、関係符号意味素（RSG)、特殊符号意味素（SSG）の類似度の計算も式（３）に従う。各パラメーターの値は

$$\alpha=1.6,$$
$$\beta_1=0.6,\beta_2=0.2,\beta_3=0.12,\beta_4=0.05,\beta_5=0.03,$$
$$\delta=0.2$$

と設定した。

４、意味解析モデル評価実験

４．１実験と評価

拡張した information structure を基にした複合語構成規則による複合名詞の意味解析モデルを評価する実験を行った。

単語数延べ 10985 件の中、部分的構文解析によって 3691 件が合成した（そのうち、人手で 1805 件の NP が判定)。ルールのマッチだけでは、意味的に膨大な曖昧性が生じるため、次に類似度の閾値を 0~1 の間に設定し、マッチする NP の数を確認してみた。比較結果は表１に示す。

類似度閾値	マッチ数	NP 合成		意味クラス推定	
		適格数	精度	適格数	精度
0.4	2602	1458	0.560338	1131	0.77572017
0.5	2069	1134	0.548091	890	0.78483245
0.6	1835	1040	0.566757	835	0.80288462
0.7	1371	808	0.589351	684	0.84653465
0.8	876	579	0.660959	518	0.89464594
0.85	444	323	0.727477	298	0.92260062

表１　類似度の閾値により合成した NP 数と推定した意味クラス

４．２誤り分析

誤りの内訳は以下のとおりである。

１．規則の不足による誤り

中国語の複合名詞の構成要素にもなれる「的」(所有、所在、時間、分量、形状、性質、原因、内容などの関係を表し、日本語の「の」に相当)、「和」(並列関係)、句読点の「、」(並列関係を表す機能もある)はほとんど今回のルールの中に組み込まなかったので、大量の複合名詞を合成することができなかった。例えば、

「中国/ns　的/u　改革/vn　开放/vn　和/c　现代化/vn　建设/vn」

「美国/ns　、/w　俄罗斯/ns　、/w　法国/ns　、/w　日本/ns　等/u」

などからは合成することができなかった。

２．ルール間の近似性により、類似度が高くても多数の意味構造とマッチし、唯一のルールに特定できなかった。

例えば、「新世纪」は同じ類似度（0.9850015）で異なるルールの用例「那-时候」「好-时候」とマッチした。

(SYN_S:TNP=A<--NSEM_S:(属性值,特)[限定]<—(时间,特,时/年/月/日)(新 ASEM_S:(属性值,特)[限定]HC:C)(世纪 NSEM_S:(时间,特,时/年/月/日)HC:H))EX:那-时候

(SYN_S:TNP=A<--NSEM_S:(属性值)[修饰]<--(时间)(新 ASEM_S:(属性值)[修饰]HC:C)(世纪 NSEM_S:(时间)HC:H))EX:好-时候

「新世紀」との類似度は同じであるが、「那-时候」「好-时候」の所属ルールが異なるところは依存関係を表す[限定]と [修饰]である。いかに比較するか更なる検討が必要である。

３．今回、人名、地名、時間、数量を表す名詞も全部一般名詞とみなしたため、数量を表す NP に時間の意味構造を付与された用例も見られる。

例えば、「许多国家」との類似度は0.96で「一-日」の意味構造が付与された。

(SYN_S:TNP=NUM<—NSEM_S:(数量值/属性值,特/属性值,时间)[限定]<—(时间,特,日)(许多 NUMSEM_S:(数量值/属性值,特/属性值,时间)[限定]HC:C)(国家 NSEM_S:(时间,特,日)HC:H))EX:一-日

５　終わりに

本稿では、information structure に基づく意味解析モデルについて述べた。まず、HowNet の information structure を出発点とし、複合語の構成規則の特徴に関する考察と拡張を施した。そして、拡張後の information structure を用いた意味解析実験、及び誤り分析を通じ、information structure によって統語構造、依存構造及び意味構造をある程度正確にとらえられることを示した。

今回は主に複合名詞 NP に対して意味解析を行った。また、中国語の単語、文の構造は実際同じ統語構造に基づくものなので、提案した手法は、単語レベルだけではなく、句や文レベルまで拡張できる。今後は複合名詞 NP 以外の複合形容詞 AP、複合動詞 VP 及び文に対しても本手法を適用し、意味解析処理の確立を試みる。

参考文献

[1] 董振東、董強（1999）「HowNet」http://www.keenage.com

[2] Qun LIU , Sujian LI “Word Similarity Computing based on HowNet,” Computational Linguistics and Chinese Language Processing Vol. 7, No. 2 , August 2002, pp. 59-76

[3] 長尾真（編）, 自然言語処理, 岩波書店, 1996

[4] Christopher D.Manning and Hinrich Schutze, Foundations of Statistical Natural Language Processing, Published May 1999 by The MIT Press Cambridge, Massachusetts