2G1-OS3-3

自然言語処理における意味研究

Research on Meaning in Natural Language Processing

松本 裕治
Yuji Matsumoto

奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

Natural Language Processing has shown a great pregress in POS tagging and syntactic parsing, making these techniques as a practical level. While semantic processing still has lots of room for improvement, various research topics have been undertaken. This article briefly overviews recent research on language meaning and tries to give future perspective.

1. はじめに

自然言語処理の分野において，形態素解析や統語解析などの言語解析技術については，90 年代以来のコーパスに基づく統計的言語処理技術により飛躍的な進歩を遂げ，精度的にも効率的にも実用レベルに達するに至ったと言える．これに対して，言語の意味に関する研究は，実用レベルとは言えないにしても，着実な進展を示している．本稿では，自然言語処理における意味研究について，最近の動向と将来について述べる．

2. 意味に関する問題点

意味とは何かということが簡単に定義できないことが意味研究の最も大きな問題である．意味を定義，あるいは，記述するための粒度が 2 つの側面から問題となる．例えば，典型的な意味処理のタスクである語義曖昧性解消を考えてみよう．ある語がいくつの語義を持つか，その粒度を決めることは容易なことではない．また，独立な単語の語義だけが意味を考える対象ではなく，複合語や複合表現など，まとまった表現を対象に意味を考える必要があることも多い．

意味に関するもう一つの問題は，意味処理への要求が応用によって異なるということである．例えば，機械翻訳では，訳語選択に必要な意味分類以上のものは求められないかもしれない．格フレームの記述や選択制約を用いた統語解析のためには，統語的曖昧性を解消するための最低限の意味分類が求められる．近年の大規模文書データを対象にした質問応答では，意味の分類よりも，表現の類似性や同義性を判定する仕組みが求められる．既存のシソーラスがどのタスクにも有効に利用できるとは限らないのと同様に，どのような目的にも使える意味処理を求めることは難しく，それぞれの目的に特化した意味研究へと分化していると言える．そのことが，言語処理における意味処理研究をわかりにくくしているのかも知れない．

3. 意味処理のタスク

典型的な意味処理のタスクとして，語義曖昧性解消 (Word Sense Disambiguation) がある．各語について，いくつの異なる語義を持つかわかっているとすると，その語の個々の使用例がどの語義で用いられているかを識別するタスクとして定義される．一つずつの単語に語義のセットを決めるのではなく，人名，地名，組織名など固有表現 (Named Entity) と呼ばれるクラスに属する語や表現を文章中から識別するタスクや，専門分野の文書から特定の意味クラスに属する専門用語を識別するタスクのように，分野や応用目的に有用な意味分類が広く行われている．

意味処理研究は，共通に利用可能なタグ付きコーパスなどの言語資源や，そのような資源を提供するコンペティション形式の会議に先導され発展したものが多い．語義曖昧性解消は，Senseval というワークショップの中心的なタスクであるし，固有表現認識は，Message Understanding Conference(MUC) の重要な部分タスクとして取り上げられた．

文の基本構造は，動詞等の用言に主語や目的語などの意味的に必須の項と場所や時間などの付加語が結合したものであり，述語-項構造 (Predicate-Argument Structure) と呼ばれる．Penn Treebank のすべての動詞に対して，同一文内に存在する項の情報をアノテーションした PropBank は，述語-項構造解析の研究を飛躍的に発展させた．近年は，同コーパス内で，事象を表す名詞に対して，やはり同一文中に現われる項の情報をアノテーションした NomBank が公開されている．事象を表す述語に対し，その項となる名詞はそれぞれ特定の意味役割をもっており，それゆえ，述語-項構造解析は，それぞれの項の役割まで同定するタスクとして，意味役割ラベリング (Semantic Role Labeling) とも呼ばれる．また，文章中の事象や時間表現の間の時間的前後関係を付与した TimeBank が公開されており，時間関係解析に役立っている．

NIST が主催する Automatic Content Extraction (ACE) Evaluation では，主たるタスクとして，実体 (Entity)，関係 (Relation)，事象 (Event)，時間表現 (Temporal Expression) などの検出と認識を共通タスクとして取り上げられている．ACE では，文書内の個々の言及 (mention) とそれが指す実体や事象を区別し，共参照関係を考慮した仕様になっている．実体は，人名や地名などの固有表現に対応し，事象は上で述べた述語となる動詞や名詞に対応する．「関係」は，実体間の様々な意味関係 (組織と構成員，製品と所有者，部分全体関係など) を意味する．

現在では NIST の Text Analysis Conference の Track になっている Recognizing Textual Entailment (RTE) Challenge では，“Text(T)” と “Hypothesis(H)” と呼ばれる 2 つの文が与えられたとし，“T” から “H” の文の意味が含意されるかどうか (含意する，T と H が矛盾する，T から H を導くことができるか不明の三者択一) を判断するという意味的推論を課題と

連絡先: matsu@is.naist.jp

している．このような推論が，質問応答，情報検索，多文書要約などの文書処理応用に必須の技術だというのがその趣旨である．因果関係などの意味的な関係をもつ事象の対を獲得する研究はホットな話題である．

文で表現される言明についての筆者の態度，特に，肯定的か否定的か，を解析するタスクは，sentiment analysis と呼ばれる．当初は文書に対して肯定的か否定的かを判定するタスクとして登場したが，近年では，文あるいはその一部を解析の対象とし，ある実体 (のある側面) に対する筆者の主観的な評価を抽出するタスクとして扱われ，opinion mining とも呼ばれる．

意味に関する基本的なタスクとして，シソーラスの (半) 自動構築，および，コーパス中の共起に基づく単語間の類似度計算，クラスタリングなどの研究が盛んに行われている．

4. 意味情報の獲得手法

前節で述べた意味処理のほとんどは，ある程度の規模のタグ付きコーパスを構築し，機械学習手法を適用するという流れで解析が行われる．しかし，前節の最後で触れた単語間の類似度計算など，大規模な未解析コーパスを用いた意味情報獲得の研究が数多くある．実応用で用いられる語彙のサイズや分野適応の問題を考えると，タグ付きコーパスからの学習ではなく，大規模な未解析コーパスの利用が重要である．本節では，大規模コーパス (未解析，あるいは，形態素解析や統語解析を自動的に行ったもの) からの意味情報獲得の代表的な手法について述べる．

2 つの語が同じ文書に近接して現れる時，それらは何からの意味的な関連を持つ (例えば,「医者」と「病院」) が，必ずしも意味的に類似しているとは言えない．2 つの語が出現する文脈 (例えば,「飲む」という動詞の目的語) が共通の場合，それらは意味的に類似していると言える．様々な文脈での出現の頻度をベクトルによって表わすと，ベクトル間の近さによって単語間の類似度を定義できる．このようなベクトルは各単語の文脈の分布を表しており，これに基づく類似度を分布類似度 (distributional similarity) と呼び，様々な尺度が定義されている (類似度については [Lin 98] の議論が参考になる)．文章中での共起を 1 次の共起，同一文脈中での出現を 2 次の共起と呼ぶと，分布類似度は 2 次の共起に基づく類似度定義である．

意味的に似ているが，たまたまコーパスの中で同一文脈に現れなかった語の対については類似度が定義できない．2 次以上の共起の情報を用いる手法がいくつか存在する．グラフに基づく手法では，まず，上記の分布類似度のような 2 単語間の類似度に基づいて単語を節点とするグラフを構成し，グラフ内の拡散やランダムウォーク等に基づいた節点間の関係を定義することにより，直接類似度が与えられていない (枝をもたない) 単語間の類似度を定義することができる．文書と単語の生成モデルとして，隠れクラスを仮定する Probabilistic Latent Semantic Analysis(PLSA)[Hofmann 99] や Latent Dirichlet Allocation(LDA)[Bei 03] というモデルがあるが，これを単語間の共起に利用することにより，各単語の隠れクラス (意味クラス) への所属確率を求めるのに用いることができる．これらのモデルでは，隠れクラスの数は恣意的に決めなければならないが，任意の単語を隠れクラスとして用いることにより潜在的な意味クラスを仮定しない Latent Words Language Model(LWLM)[Deschacht 09] というモデルが提案されている．

意味的な類似度計算のための単語共起には，前後数単語の窓内の共起や動詞とその目的語としての共起など，特定の構造における共起が用いられる．一方，ある特定の文脈やパターンが，そこに出現する実体の意味や実体間の関係を強く規定する場合がある．Hearst[Hearst 92] が実体間の IS-A 関係を抽出するために，そのような関係を持つ単語が出現する典型的な言語表現パターンを定義することにより，大規模コーパスから自動的に上位-下位関係にある語を抽出した研究が有名である．特定の実体や実体間の関係を獲得するために個々にパターンを記述することは大変な労力を要するし精度やカバレージも保証されない．少ない事例をシードとして用意し，パターンの獲得と事例の獲得をそれぞれの信頼度を考慮しながらブートストラップ的に行う Espresso というアルゴリズムを Pantel らが提案している [Pantel 06].

5. おわりに：意味研究の将来

最初に述べたように，自然言語処理研究では，統語解析までの表面的な言語解析は飛躍的な進歩を遂げ，大規模な実データを高速かつ高精度で解析することが可能になってきた．しかし，ほとんどの応用分野では，必ずしも単語の品詞や統語構造などが必要なわけではない．これらの解析を行う理由は，これらの情報が応用分野で求められている問題解決の精度向上に有用と考えているからである．本当に求められているのは，述語-項構造解析や事象間の因果関係や時間関係，表現の同義性のような，意味に直接関係する情報なのであろう．意味情報抽出の精度に直接影響する言語解析とは何かを考える必要があるだろう．

90 年代までには，単一化文法，語彙概念構造，生成語彙など，統語と意味を結びつける語彙化文法や語彙意味論と言われる分野が進展したが，2000 年代にはこの方面での目立った進歩が見られない．格フレームや語義ラベルなどの表層的な情報，ましてや類似度のような情報ではなく，語の意味を記述するためのより深い構造とその上での演算と，統計的な情報を融合したより豊かな意味記述の枠組みの進展に期待したいと思うが，著者の好みに偏りすぎであろうか．

なお，紙面の関係で近年の意味処理に関連する様々な研究や文献を挙げることができなかった．また，本稿で触れたコーパスやプロジェクトは Web で簡単に発見できるので，あえて URL は掲載しなかった．

最後に，日頃より様々な情報提供をいただく同僚，学生諸氏に感謝したい．

参考文献

[Bei 03] Blei, D.M. et al., ”Latent Dirichlet Allocation” Journal of Machine LLearning Research, vol.3, (2003).

[Deschacht 09] Deschacht, K. and Moens, M-F., ”The Latent Words Language Model,” Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning, (2009).

[Hearst 92] Hearst, M., “Automatic Acquisition of Hyponyms from Large Text Corpora,” 14th International Conference on Computational Linguistics, (1992).

[Hofmann 99] Hofmann, T., ”Probabilistic Latent Semantic Analysis”, Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, (1999).

[Lin 98] Lin, D., “An Information-Theoretic Definition of Similarity,” Proceedings of International Conference on Machine Learning, (1998).

[Pantel 06] Pantel, P. and Pennacchiotti, M., “Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations,” COLING/ACL-06, (2006).