1P−1

自然言語処理のための知識辞書および知識推論部について

松平 正樹　網島 督之　坂本 仁
沖電気工業(株)

0.はじめに

機械翻訳の訳質を向上するために共起などの知識を活用することが有効であるという報告が多数なされているが([1])、実際の文章での応用例や検証がなく、効果が不明確であった。

そこで、共起データを知識として格納する知識辞書、および、知識辞書に存在する知識から必要な知識を推論する知識推論部を作成し、実際の文章から共起データを抽出して訳質の向上に有効であるかを検証した。今回は、コンピュータ関連の文章の並列名詞句の曖昧性の解消について実験を行なった。

1章で知識辞書と知識推論部、2章で実際の文章による実験方法および実験結果について述べる。最後に3章でまとめと今後の課題について述べる。

1.知識辞書と知識推論部

知識辞書は、実際の文章から抽出した共起データの集まりで、例えば、「コンピューターのCPU」、「ネットワークに接続する」などの形で表現される。実際には、共起する2つの単語、それらを接続する単語、それぞれの単語に対する品詞、共起の確信度の4項目から構成される。知識の確信度は以下の式によって求める([2])。

$$Wix = 2/(1+\exp(W'ix))-1$$

$$W'ix = (1-\exp(-Ci*C'x/T))*(Cix-Ci*C'x/T)/(Ci*C'x/T)^{1/2}$$

Ci：単語iが最初の構成単語となる共起データの個数

C'x：接続語+単語xが構成単語となる共起データの個数

Cix：単語iと接続語+単語xが構成単語となる共起データの個数

T：単語+接続語+単語のパターンの共起データの総数

知識推論部は、意味距離計算部、意味距離格納部、知識生成部より構成される。

意味距離計算部は、あらかじめ以下の式を用いて共起データから単語間の意味的な距離を計算し、意味距離格納部に格納する。

$$Dij = 1-\Sigma_k(Wik*Wjk)/(1+\Sigma_k((1-\exp(-Ci*C'k/T))*(1-\exp(-Cj*C'k/T)))^{1/2})$$
$$(k:Wik > 0 \text{ or } Wjk > 0)$$

知識生成部は、知識辞書に格納されている知識と意味距離格納部に格納されている情報から新しい知識を推論する。例えば、「AのB」、「XがYする」からAとXが意味的に近ければ、「XのB」、「AがYする」という知識を推論する。また、元の知識の確信度と単語間の意味的な距離から推論された知識の確信度を計算する。

2.実際の文章による実験

コンピュータ関連の文章約25万文(対象文)を用いて、並列名詞句の曖昧性の解消の実験をおこなった。

2.1 並列名詞句の曖昧性の解消方式

入力文「AとBのC」に対して、AとBが並列/AとCが並列の2つの可能性がある。この曖昧性を知識を用いて次のように解消する。

(1)「AとB」あるいは「AのC」という知識が存在すればAとBが並列であると判断する。

(2)「AとC」という知識が存在すればAとCが並列であると判断する。

2.2 知識の獲得

知識辞書の知識として、対象文から名詞+「の」+名詞、名詞+「と」+名詞のパターンの共起データを抽出し確信度を計算した。以下に抽出した共起データの総数と種類数を示す。

	総数	種類数
名詞+「の」+名詞	89,380	16,386
名詞+「と」+名詞	6,364	2,751

2.3 意味的な距離の計算

頻度が高い上位1000語の名詞について対象文から名詞+助詞+サ変動詞のパターンの共起データを抽出し、単語間の意味的な距離を計算した。

2.4 並列名詞句の抽出

対象文の一部(全体の約4%)から曖昧性のある並列名詞句(名詞+「と」+名詞+「の」+名詞のパターン(AとBのC))を抽出し、特別な単語により自明なもの(AとBの間、Aと同様のCなど)を削除し、実験の対象とした。また、辞書引きをおこない既存の意味情報の大分類(人間、組織、場所など20種類)を付与した。対象となる並列名詞句は81件であった。その一部を

Knowledge Dictionary and Knowledge Reasoning for Natural Language Processing
Masaki MATSUDAIRA, Tadayuki TSUNASHIMA, Masashi SAKAMOTO
OKI Electric Industry Co., Ltd.

以下に示す。

CADとCAMのユーティリティ
アプリケーションとサービスの必要性
オフコンとパソコンの接続
パソコンと会社のメインフレーム
新製品とサービスベンダーの出現
端末とホストのインターフェース

2.5 実験の方式および結果

まず、「AとBのC」に対して人間が判断して正解を作成した。その個数を以下に示す。

AとBが並列	63	77.8%
AとCが並列	18	22.2%

次に、2つの方式で曖昧性の解消の実験を行なった。

方式1(従来の意味情報)
(1)Aの意味情報=Bの意味情報、かつ、Aの意味情報≠Cの意味情報ならば、AとBが並列と判断する
(2)Aの意味情報≠Bの意味情報、かつ、Aの意味情報=Cの意味情報ならば、AとCが並列と判断する

方式2(知識辞書+知識推論)
(1)知識辞書に「AとB」が存在する場合、AとBが並列
(2)知識辞書に「AとC」が存在する場合、AとCが並列
(3)知識辞書に「AのC」が存在する場合、AとBが並列
(4)あるXに対し、AとXが意味的に近く、知識辞書に「XとB」が存在する場合、AとBが並列
(5)あるXに対し、BとXが意味的に近く、知識辞書に「AとX」が存在する場合、AとBが並列
(6)あるXに対し、AとXが意味的に近く、知識辞書に「XとC」が存在する場合、AとCが並列
(7)あるXに対し、CとXが意味的に近く、知識辞書に「AとX」が存在する場合、AとCが並列
(8)あるXに対し、AとXが意味的に近く、知識辞書に「XのC」が存在する場合、AとBが並列
(9)あるXに対し、CとXが意味的に近く、知識辞書に「AのX」が存在する場合、AとBが並列
と判断する
ただし、上記(4)〜(9)で「意味的に近い」とは、単語間の意味的な距離が1以下とする。

結果を以下に示す。

	決定数(正解数)	決定率(正解率)
方式1	45(39)	55.6(86.7)
方式2	46(44)	56.8(95.7)
(1)〜(3)	27(27)	33.3(100.0)
(4)〜(9)	19(17)	23.4(89.5)

また、方式2で用いた推論の例を以下に示す。

・互換機の出現
＋新製品≈互換機
—› 新製品の出現
・パソコンのインターフェース
＋端末≈パソコン
—› 端末のインターフェース

結果から、方式1は決定率、正解率とも低く、方式2は決定率は十分ではないが正解率が高いことがわかる。
そこで、方式1と方式2を統合した以下の方式で再び実験を行なった。

方式3(知識辞書+知識推論+従来の意味情報)
(1)方式2を適用する
(2)方式1を適用する

結果を以下に示す。

	決定数(正解数)	決定率(正解率)
方式3	71(68)	87.7(95.8)
(1)	46(44)	56.8(95.7)
(2)	25(24)	30.9(96.0)

結果から、方式3の方法は決定率、正解率とも高く、並列名詞句の曖昧性の解消に非常に有効であることがわかる。また、方式2、方式3の各段階での決定数から、知識辞書、統計的に計算した単語間の意味的な距離からの推論、従来の人手で付与した意味情報がすべて曖昧性の解消に寄与していることがわかる。これは、知識辞書そのままではヒット率が低く、単語間の意味的な距離からの推論では頻度の低い単語について情報が少なく、従来の意味情報では未知語や分野特有の意味(語用)に対応できないといったそれぞれの欠点を補い合っているためと思われる。

3.まとめ

実際の文章から共起データを抽出して知識辞書および知識推論部を作成し、それらを用いて並列名詞句の曖昧性の解消の実験を行なった。実験の結果、知識辞書、知識推論部を従来の意味情報と組み合わせる方式が、並列名詞句の曖昧性の解消に非常に有効であるという結果を得た。
今後は、修飾先の決定、訳語選択などについて知識辞書および知識推論部の有用性を検証していく予定である。

参考文献

[1]:田中他:「自然言語の知識獲得」:情報処理学会自然言語処理 65-2 (1988)
[2]:松平:「共起データを用いた単語の意味ネットワークの作成」:情報処理学会第42回全国大会 (1991)