高次 N-gram を用いた形態素解析の研究*

村上仁一 (NTT 情報通信研究所)

1 まえがき

近年、N-gram を用いた形態素解析の研究が盛んになっている [1]。しかし、これらの研究では、次数が 2(bigram) もしくは 3(trigram) であり、高次での結果はあまり報告されていない。本論文では、漢字仮名の 6-gram を使用したときの形態素解析の実験結果を報告する。

ここでは、検索のために企業名に単語区切りを入れる目的で形態素解析を行なった。まず電話帳から全国の企業 440 万件を選出し、人手によって単語区切りをいれた。次にこのデータから単語辞書と漢字仮名の N-gram の連鎖確率値を計算した。最後に、これらのデータを用いて新規 1 万件の企業名に対して形態素解析を行い、人手によって単語区切りを入れた結果と比較した。この結果 6-gram を使用することで 1 位正解率で 83.6% 2 位正解率で 89.2%の正解率が得られた。

2 形態素解析

図 1に一般的な形態素解析のブロック図を示す。

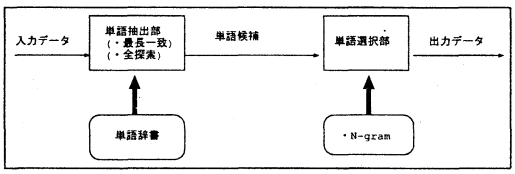

図 1: 形態素解析のブロック図

形態素解析には多くの目的がある。本論文では、企業名のキーワード検索を目的とした形態素解析を考えた。キーワードを使用して検索する場合、企業名を形態素解析して単語区切りをいれることにより、検索精度が向上することが知られている。

3 実験条件

3.1 形態素解析データベース

実験データとして電話帳の企業名を使用した。電話帳から企業数約 440 万件を選出し、人手によって形態素解析を行なった (以下、形態素解析データベース)。このデータベースを学習データとして単語辞書および N-gram の連鎖確率値の計算に使用した。また、新規の企業名 1 万件を電話帳から抜き出し、これを test データとした。

表 1に形態素解析データベースの例を示す。

表 1: 形態素解析データベースの例

企業名 (入力データ)	形態素解析データ (出力データ)
あさひだるま	あさひ+だるま
お好み焼童子	お好み焼%童子
くるまやラーメン	くるま|や+ラーメン
ろばた焼童子	ろばた焼%童子
江南赤童子店	江南%赤童子|店

本データベースには、3種類の単語区切りが使用されている。これらの単語区切りの意味を以下に示す。

\| : 接辞境界

接辞境界の前後は接頭語もしくは接尾語になる。(例 : や、店)

\+ : 単語境界

本論文ではキーワード検索を目的としたため、本データベースの単語は、通常の形態素解析の単語より短い傾向にある。(例: あさひ、だるま)

*" A Study of High order N-gram model " by Jin'ichi Murakami (NTT Information and Communication System Laboratories)

% : アクセント句境界

本データベースは、合成音声で企業名を出力できるようにするため、人間がポーズをつける単語境界をアクセント句境界とした [2]。

3.2 単語抽出部

単語抽出部のアルゴリズムとして、最長一致法や文節数最小法などが良く使用されている。しかし、最近のコンピュータのコストの低下に伴い、全ての候補を計算する全探索法 (tree-trellis search) も可能になってきた。本論文では、最長一致法と全探索法で実験を行った。

3.3 単語辞書

単語辞書には、学習データ約 440 万の企業名を、単語区切り（アクセント句境界と単語境界と接辞境界）ごとに分割し、これらを単語辞書として登録した。例えば “くるま | や+ラーメン” は “くるま |”、“| や+”、“+ラーメン” の 3 単語を単語辞書に登録する。

3.4 単語選択部

本論文では、単語抽出部で出力された複数の候補選択するために、漢字仮名の N-gram を用いた。また、アクセント句境界および単語境界および接辞境界は 1 単語として計算した。

例えば 2-gram(bigram) では、“くるま | や+ラーメン” の連鎖確率値は

$P(\text{く}/start) \times P(\text{る}/\text{く}) \times P(\text{ま}/\text{る}) \times P(|/\text{ま}) \times P(\text{や}/|) \times P(+/\text{や}) \times P(\text{ラ}/+) \times P(\text{ー}/\text{ラ}) \times P(\text{メ}/\text{ー}) \times P(\text{ン}/\text{メ}) \times P(end/\text{ン})$

と計算した。

実験では N の次数を 2(bigram) から 6 まで変化させて正解率を調査した。

4 実験結果

実験は人間によって与えられた形態素解析結果と完全に一致する候補を正解候補として計算した。実験結果を表 2に示す。横軸は N-gram の次数で縦軸は正解率である。また 4 位までの累積正解率も示した。

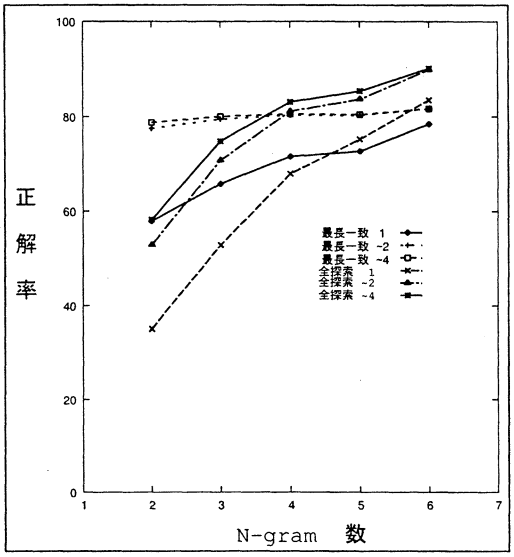

図 2: 実験結果

この結果から N-gram の次数をあげるに従い、正解率が向上することが示された。そして、全探索法で 6-gram を使用したとき 1 位正解率で 83% 2 位正解率で 90%の高い精度が得られることが示された。

5 考察

5.1 N-gram の次数

本実験では N-gram の次数が上がるほと正解率が向上することが示された。これは、実験データが非常に限られた分野（企業名）であるためと考えられる。今後、次数をあげて正解率がどこまで向上するか調査する予定である。

5.2 正解率

本実験では、人間によって与えられた形態素解析結果と完全に一致する候補を正解として計算した。しかし、アクセント句境界と単語境界は特に曖昧である。そのため、正解と見なせる候補を誤りとしている例も多い。そのため、全探索法において 6-gram を使用した実験結果において、単語区切りの場所が同じで種類が異なる件数を調査した。この数は 793 件あった。この例を表 2に示す。これらを正解にす

表 2: 実験結果（単語区切りの記号の異なる例）

人間による形態素解析結果	1 位候補
日本橋%大伝馬 | 町%郵便 | 局	日本橋%大伝馬 | 町+郵便 | 局
日本+経済+新聞%長瀬+販売 | 所	日本%経済+新聞%長瀬+販売 | 所
能代%国道+維持+出張 | 所	能代%国道+維持%出張 | 所
萩原+カイロプラクティック	萩原%カイロプラクティック
白峰 | 村%公民 | 館%事務 | 所	白峰 | 村+公民 | 館%事務 | 所
八百松 | 亭	八百松+亭
飯坂 | 新+会館	飯坂+新%会館
美瑛 | 川%砂利+砕石%販売+協業+組合	美瑛 | 川%砂利%砕石+販売%協業+組合
浜田 | 屋%食料 | 品 | 店	浜田 | 屋+食料 | 品 | 店
富山 | 県%鍼灸+マッサ | ージ | 師 | 会	富山 | 県%鍼灸%マッサ | ージ | 師 | 会
富士宮%ホワイト%テニス+クラブ	富士宮%ホワイト+テニス+クラブ
部落%解放+同盟%鹿児島 | 県%連合 | 会	部落%解放+同盟%鹿児島 | 県+連合 | 会

表 3: 実験結果（単語区切りの場所が異なる例）

人間による形態素解析結果	1 位候補
神結+酒造	神%結%酒造
神緑+薬局	神%緑%薬局
紳士服 | の%高村	紳士 | 服 | の%高村
諏訪山+公園%管理+事務 | 所	諏訪 | 山+公園%管理+事務 | 所
杉の沢	杉 | の+沢
世海	世%海
瀬古勝+製菓 | 舗	瀬古%勝+製菓 | 舗
瀬川%米穀+酒類 | 店	瀬川%米穀+酒 | 類 | 店
栖来 | 寺	栖%来%寺

ると 1 位正解率は 91%になる。

また、単語区切りの場所が違うものは 944 件あった。この例を表 3 に示す。この結果をみると、人間による形態素解析結果が誤っていると思われる例もある。これらの例を除くと実際の正解率はかなり高いと思われる。

6 まとめ

本論文では、検索のために企業名に単語区切りを入れる目的で、漢字仮名の 6-gram を使用したときの形態素解析の実験結果を報告した。まず電話帳から全国の企業 440 万件を選出し、人手によって単語区切りをいれた。このデータから単語辞書と漢字仮名の N-gram の連鎖確率値を計算した。そして、これらのデータを用いて新規 1 万件の企業名に対して形態素解析を行い、人手によって単語区切りを入れた結果と比較した。この結果 6-gram を使用することで 1 位正解率で 83.6% 2 位正解率で 89.2%の高い正解率が得られた。

また、誤りとされた候補の中にも正解と見なせる候補が多く、実際の 1 位正解率は 91%を越えることが示された。

参考文献

[1] 北 , 中村, 永田, “音声言語処理”, 森北出版.

[2] 東田 他, “オペレータレス自動電話番号検査システムの開発”, 自然言語処理研究会 98-NL-123-4 pp.25-32 (Jan. 1998).