E-017

ウイグル語からの自動辞書構築に関する研究

Study on automatic dictionary construction from Uighur language

マヒヌリズパル†、吉田稔†、松本和幸†、北研二†
Mayinuer Zipaer, Yoshida Minori, Matumoto Kazuyuki, Kita Kenji

1. はじめに

類似語の同定は,情報検索,テキストマイニングなどのテキスト処理を行う上で必要な作業である、同義語辞書を作成することとテキスト処理の効率や精度の向上を期待できる.

本研究ではウイグル語辞書の自動構築をするため、ウイグル語ニュースサイトの記事中の単語を word2vec でベクトル化して、類似語のリストを作成した。

1.1 ウイグル語

ウイグル語は中国領内の話者人口は 1100 万人以上で、ウズベク語、カザフ語などが含まれるチュルク語族の中でである一つ言語である。

ウイグル語は日本語と同じく膠着語に分類されており、語幹に接頭辞、接尾辞が膠着されるによって語が形成される。アラビア式のウイグル語文字を正書法で使用しており、文が空白にょって分かち書きされる。分かち書きされた単位は形態素と同じではないことが多い。今まで存在してるウイグル語処理システムでは、システム毎に語の認定が異なっている。空白で空間が切られた文字例を一つ語とすることもあり、これで、辞書に接辞を語幹に付けた形で登録すれば、辞書の解析精度は高くなるものの、辞書のサイズが急激に大きくなる。語幹を中心で語を認定することもあり、この方法では、第一の語幹を自立語として、他の部分を接辞として記述する。

1.2 関連研究

ウイグル語自動辞書構築する研究は、今まで数多く行われってきた。小川らは、実用に近い日本語-ウイグル語機械翻訳システムの実現を目指して一連の研究をしてきている[2,3,4,5,]。彼らは、一定の語彙数を持つ日本語-ウイグル語電子辞書の開発の不可欠であると考え、まず IPA の計算機用日本語基本動詞辞書 IPAL をベースに、名詞や形容詞等を含め、約 1200 語の日本語-ウイグル語電子辞書を作成した。さらに、この 1200 語の辞書を、日常使われる最低限の語彙を含むよう拡張した。具体的には、まずウイグル語辞典を電子化し、16000 語のウイグル語-日本語電子辞書を作成した、その逆辞書を半自動作成して、これにより、語数約 2 万の実用に近い日本語-ウイグル語電子辞書を作成した。また、日本語-ウズベク語の機械翻訳システムを開発したほか、ウイグル語とウズベク語の語彙の類似度を利用して、日本語-ウズベク語辞書の拡張も行っている。

2. 提案方法

本研究はウイグル語辞書を自動構築することために、類似語リストを作成することを目標する。本研究では、word2vec[8]を使用してワードベクトル化し、そのベクトルの類似度を利用することで、ウイグル語の各単語に対して類似する単語のリストを取得する。得られたリストが実際に類似語となっているか否かを、様々な単語に対する類似語リストを取得することで調査する。

2.1 Word2vec

Word2vec は Mikolov らによって提案された、ニューラルネットワークを用いて単語の分散意味表現を計算する手法である。分散表現とは、単語を実数ベクトルで表現することである。同じ文脈で出現する単は類似した意味を持つという分布仮説に基づき、ある単語例を与えられた時、次に出現する単語を予測するというタスクをニューラルネットワークに学習させ、文脈を考慮した単語の分散表現を得る。本研究では word2vec を用いることでウイグル語ニュース記事からコーパスを作成して、学習を行う、Python パッケージの gensim 内に実装されている word2vec を用いて類似語のリストを作成する。

2.2 実験

中国人民日報のネットバージョンのウイグル語バージョンから、197635（約 20 万）語のウイグル語単語のコーパスを作成した。

まずウイグル語辞書の自動構築をするため、ウイグル語ニュースサイトの記事中の名詞、動詞、数量を word2vec でベクトル化し、類似語のリストを作成した。各単語に対して、類似語リストのサイズを 10 語,20 語,30 語,…,100 語それぞれとした場合の結果を調査した。

2.3 実験結果

以下に、リストサイズを 10 語と設定した時の結果例を示す。類似語と判定できる語の率は 91%となった。

رۇسىيە	Russia	ئۈچ	Three
○ ئۇكرائنا	Ukraine	○ تۆت	Four
▲ روسيەنىڭ	Russian	○ ئالتە	Six
○ ئىران	Iran	○ بەش	Five
▲ رۇسىيەنىڭ	Russian	○ ئىككى	Two
○ قازاقستان	Kazakhstan	○سەككز	Eight
○ ئۆزبېكستان	Uzbekistan	○ ئۈچ	Three
○ تاجىكستان	Tajikistan	○ يەتتە	Seven
○ تۈركمەنستان	Turkmenistan	○ توققۇز	Nine
○ ئۇكرائنيە	Ukraine	○ بىرقانچە	A few
○ ھىندىستان	India	○ بىرنەچچە	A few
100%		100%	

表 1

†徳島大学, Tokushima University

本研究で関連語も類似語として判定されている。（表 1 に示す。）類似語を○で、関連語を▲でマークした。ئۇروسىيەن のような、入力単語(青色単語)の語幹に接尾辞が付いている語形も関連語と判定している。（単語の赤色部分は接尾辞である。）

分類された類似語を 10,20,30,…,100 単語リストまで設定して実行した時の結果は平均正解率 60%である。（図 1 に示す。）

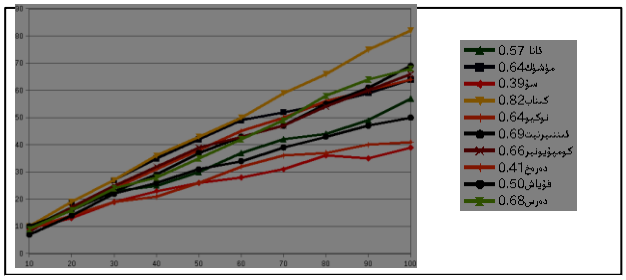

図 1　実験結果

3. おわりに

本研究では、ウイグル語からの自動辞書構築のために、ウイグル語ニュース記事から約 2 万語のコーパスを作成し、woed2vec による学習行い、類似語リスト作成した。

類似語リストのサイズを 10 単語と設定した時の類似語の率は 91%と高かった。しかし、リストサイズの増加とともに、正解率は低下した。100 単語リストの正解率（類似語率）は 60%であった。

今後の課題として、現在取得できているコーパスのサイズはまだ小さく、類似語の正解率を上げるために、これを増やす必要がある。ほかのウイグル語ネットバージョンからもウイグル語単語を大量に収集し、ウイグル語コーパス作成する。また、取得した類似語の中で、入力単語の語幹に接頭辞や接尾辞が付いた語形が多くなった。本研究ではそれも類似語として判定されているが、これを不正解と判定した場合、正解率は 10 単語リストで 91%から 36%と低下する。このような類似語への対処も今後の課題である。

どの単位までの一つ語にするか、品詞を基準するか、あるいは意味を基準するか、それぞれを実験して、ウイグル語の分かち書き単位の語を認定する。

参考文献

[1] 寺田昭,吉田稔,中川裕志,“同義語辞書作成支援ツール”，自然言語処理, Vo1.13,No.2.(2006).
[2] ムフタル　マフスット,小川 泰弘,杉野 花津江,稲垣 康善,“日本語 – ウイグル語辞書の自動作成とその収録語の分析”，自然言語処理, 151−2, No.n (200２).
[3] ムフタル　マフスット,小川 泰弘,杉野 花津江,稲垣 康善,“日本語ーウイグル語辞書の半自動作成と評価”自然言語処理,Vo1.10,No.4.(2003).
[4] 小川 泰弘,福田ムフタル,外山 勝彦“日本語対訳辞書拡張のためのウイグル語からウズベク語への翻字手法”,言語処理学会,(2008).
[5] 小川 泰弘,福田ムフタル,外山 勝彦“日本語ーウイグル語翻訳掲示板システム”,言語処理学会,(2009).
[6] https://ja.wikipedia.org/wiki/ウイグル語
[7] https://en.wikipedia.org/wiki/Uyghur_alphabets
[8] https://code.google.com/archive/p/word2vec
[9] https://radimrehurek.com/gensim/
[10] 鹿島 好央,北山 大輔“Word2Vec と Web 検索を用いた検索クエリ置換手法”,DEIM Forum 2017 C6-1・
[11] アブドレイム　アブドハリリ,伝康晴,土屋俊 “ウイグル語接辞の頻度について”, 言語処理学会.