確率的用例ベース翻訳の実現

荒牧英治 黒橋禎夫
柏岡秀紀 加藤直人

発表者：林華

はじめに

コーパスベース翻訳には

- 統計ベース翻訳
- 用例ベース翻訳

用例ベース翻訳の基本的なアイデア:

入力文の各部分に対して類似度高い用例を選択し、それらを組み合わせて翻訳を行う。

提案手法

基本的な原則：できるだけ大きなサイズの用例を用いて翻訳を生成することである。

例：“彼はCDをかける”

誤り：bet、lack、break、run...

正解：play

したがって、”CDをかける”という大きな用例のほうがいい

この手法の手順

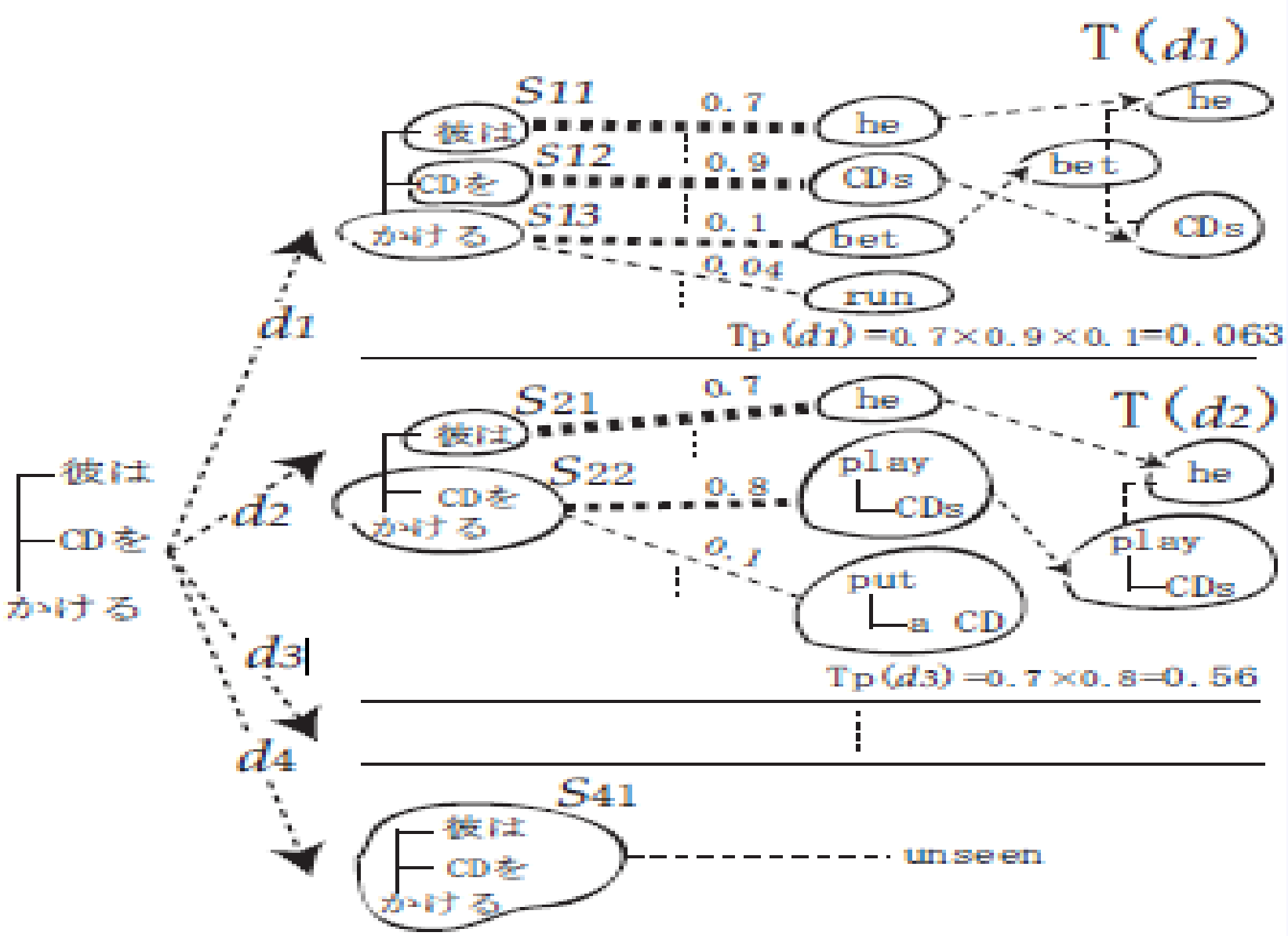
1. 入力文を可能なかぎりの部分木の組合せに分解する。 D = {d1, ..., dN}.
2. di は入力文をMi 個の部分木に分解しているとする。 di = {si1, si2, ..., siMi}
3. 各部分木sij について, もっとも翻訳確率P(tij | sij)の高い用例を選び, それらの積を翻訳文の翻訳確率P(di) とする。
4. , 最も高い翻訳確率を持つdm を以下の式によって探索し, 最終的な翻訳T(dm)を得る

上の手順の式

$$P(d_i) = \prod_{s_{ij} \in d_i} \max_{t_{ij}} P(t_{ij} \mid s_{ij})$$

$$d_m = \arg\max_{d_i \in D} P(d_i)$$

手法説明の例


T(d1)
S11
0.7
彼は
he
S12
0.9
CDを
CDs
S13
0.1
かける
bet
0.04
run
he
bet
CDs
Tp(d1)=0.7×0.9×0.1=0.063
d1
d2
d3
d4
彼は
CDを
かける
S21
0.7
彼は
he
S22
0.8
CDを
かける
play
CDs
0.1
put
a CD
T(d2)
he
play
CDs
Tp(d3)=0.7×0.8=0.56
S41
彼は
CDを
かける
unseen

パラメータの推定

まず，英語部分木t と日本語部分木s からなる用例があるとする。この翻訳確率P(t | s) は、アライメントされたコーパス中での対応(t, s) の出現頻度を直接数えて求める:

$$P(t \mid s) = \mathrm{count}(t, s)/\mathrm{count}(*, s)$$

用例を次のスライドに示す

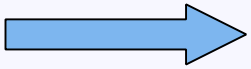
英語と日本語との部分木用例

英語側	日本語側(とコンテキスト)	context_sim
play	(CDを)かける	0.8
play	(テープを)かける	0.8
put	(MDを)かける	0.8
⋮	⋮	⋮
set	(目覚ましを)かける	0.6
⋮	⋮	⋮
bet	(お金を)かける	0.7
bet	(財産を)かける	0.7
bet	(100ドルを)かける	0.3

不具合な所

コーパスに存在しない句には、部分木が単独で翻訳確率を計算することになり、不適切な訳語が選ばれる可能性がある。

例：“レコードをかける”という用例存在しない。
よって、不適切な訳語が選ばれる可能性高い。

ここで  用例のフィルタリング行う

contextsimによる用例フィルタリング

用例Aと入力文のコンテキストの類似度を次の式で定める:

$$context\_sim(A) = \sum_{i \in N} sim(i, j)$$

i は用例A の日本語側で翻訳に使う部分の周辺の句, j はi と対応する入力文の句, N はi の集合, sim(i, j) はシソーラスを用いて計算するi とj の類似度(max=1) である.

用例フィルタリング―その２

用例A の翻訳確率P(t | s)を計算する際には、context_sim(A) 以上の類似度を持つ用例だけを集計して翻訳確率を計算する。

上の例では:

P(play | かける) = 2/3 ,

P(put | かける) = 1/3 となる.

実験

- 提案システム(proposed)と経験則によるメジャーにより用例を選択するシステム(basic) を比較する.
- 自動評価法を用いる
- コーパスはIWSLT04 にて配布されたコーパス(トレーニングとテストセット) を用いた.
- 翻訳辞書を用いた手法[4] でアライメントを行った
- テストセットは日本語文(500 文) とそれらの16 通りの英語翻訳(500×16文) からなる.

自動評価手法

BLEU　正解とのn-gram の適合率の相乗平均
NIST　正解とのn-gram の適合率の相加平均
WER Word Error Rate.　正解との編集距離
PER　Position Independent Word Error Rate. 語順を用いない正解との編集距離
GTM　general text matcher. 正解との一致した最長語列の適合率，再現率の調和平

実験結果

	bleu	nist	wer	per	gtm
PROPOSED	0.41	8.04	0.52	0.44	0.67
BASIC	0.39	7.92	0.52	0.44	0.67
WITHOUT_SIM	0.42	7.67	0.49	0.42	0.68

用例のフィルタリングの効果

実験結果は，NIST においてはproposed が高く，BLEU においてはwithout sim が僅かに高かった．NIST は訳語選択により鋭敏に反応するため，コンテキストの類似度は訳語選択に貢献すると考えられる．

入力文: このカートは使えますか.

proposed: Can I use this cart ?

without sim: Do you accept this cart ?
