

統計的機械翻訳入門

統計的機械翻訳に関する講習会

2008.8.28-29

山本幹雄 筑波大学

Patent Translation Task Group,
Cluster of Focused Domains,
NTCIR (NII Test Collection for IR Systems) Project

講習会後修正版(2008.9)

講義目的と内容

目的

- 実習で用いるシステムの仕組みを理解する
 - 特にMosesという統計的機械翻訳システムについて理解する

内容

- 統計的機械翻訳の概要
 - 言語モデル: n gramモデル
 - 翻訳モデル: フレーズ翻訳モデル
 - デコーダ: Multistack & Beam-search
- フレーズ翻訳モデルの推定
 - 方向のある単語対応 (IBMモデル, GIZA++)
 - 対称化単語対応 (講義では省略)
 - フレーズベアの抽出
- 自動翻訳性能評価
- チューニング: Minimum Error Rate Training

午後の実習の流れ

- 訓練データの準備
- 言語モデルの構築
- フレーズ翻訳モデルの構築と確認
- Moses設定ファイルの作成
- チューニング
- 翻訳実験と評価
- 最終課題

この入門よりも少し詳しいチュートリアル: Philipp Koehn, Statistical Machine Translation: the basic, the novel, and the speculative. Tutorial at EACL 2006. <http://www.aclweb.org/anthology/E06/EACL06/tutorial/koehn/>

統計的機械翻訳

- Corpus-based NLP → データ + 人間の知識
重視の度合い: 経験主義 ↑ 合理主義 ↑
- 経験主義の最も極端な立場: 統計的機械翻訳

大量の言語データ

- プレーンテキスト (新聞・WEBなど)
- 対訳テキスト (議事録・特許文など)



パラメータの自動推定 (数理統計学)

翻訳の (確率) モデル 翻訳辞書
単語の並び替え

対訳コーパスの例: Hansards

1980年代から100万文ペア以上存在 (カナダの国会議事録)

英語

The 35th Parliament having been dissolved by proclamation on Sunday, April 27, 1997, and writs having been issued and returned, a new Parliament was summoned to meet for the dispatch of business on Monday, September 22, 1997, and did accordingly meet on that day.

Monday, September 22, 1997.

This being the day on which Parliament was convoked by proclamation of His Excellency the Governor General of Canada for the dispatch of business, and the members of the House being assembled:

フランス語

La trente-cinquième législature ayant été prorogée et les Chambres dissoutes par proclamation le dimanche 27 avril 1997, puis les brefs ayant été émis et rapportés, les nouvelles Chambres ont été convoquées pour l'expédition des affaires le lundi 22 septembre 1997 et, en conséquence, se sont réunies le jour dit.

Le lundi 22 septembre 1997.

Le Parlement ayant été convoqué pour aujourd'hui, par proclamation de Son Excellence le Gouverneur général du Canada pour l'expédition des affaires, et les députés étant réunis:

<http://www.isi.edu/natural-language/download/hansard/>

対訳コーパスの例: NTCIR-7 (日米特許)

約180万文ペア

英語

The lock flag is outputted to the controller 2 from the register 6.

A plurality of clients are connected to the server.

Next, the server application program issues a recv command 2116 by designating a data receiving buffer 2161.

As shown in FIG. 2, a counter 5 and a flip-flop 6 generate a pulse having a width proportional to the value (Tf) calculated by signal processing circuit 4.

A gate electrode 61 is formed on the periphery of gate insulating film 63 in trench 53.

The process for translating the input sentence is the same as in the above-mentioned example.

日本語

また、このロックフラグは、レジスタ6からコントローラ2に出力される。

サーバには複数のクライアントが接続する。

次にサーバアプリケーションプログラムはデータ受信バッファ2161を指定してrecvコマンド2116を発行する。

図2を参照すると、カウンタ5およびフリップフロップ6は、信号処理回路4によって計算された値(Tf)に比例した幅を持つパルスが発生するためのものである。

ゲート電極61はトレンチ53内に位置し、ゲート絶縁膜63の周囲に形成されている。

入力文を翻訳する過程は、上記した例と同様である。

...

何故、いま統計的機械翻訳なのか？

- 性能の目覚ましい向上 (ここ5年)
- 性能向上の原因
 - 翻訳モデルの発展
 - IBMモデル (単語) → フレーズモデル
 - デコーダ等の仕組みが公開
 - 10年前は職人技
 - ツールの普及
 - GIZA++, SRILM, Pharaoh, Moses
 - 自動評価手法の開発と普及
 - BLEU
 - チューニング手法 (+ 識別モデル)
 - MERT: Minimum Error Rate Training
 - 対訳データの拡充
 - Hansards (英仏) → Europarl (11ヶ国語100万文以上並記)
 - 特許対訳データ (NTCIR-7): 日英180万文

ACL Best Paper Award
ACL2001: デコーダ
ACL2002: 識別モデルと
チューニング
ACL2005: 翻訳モデル

Google 機械翻訳 アラビア⇄英語
2006.5よりサービス開始

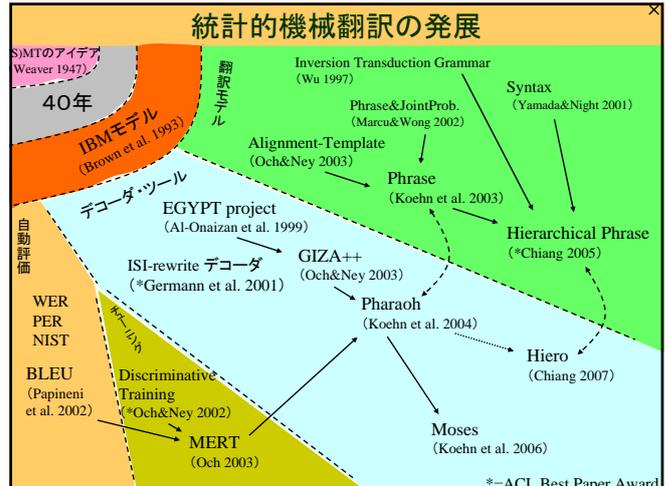
ユニセフのあるページ (NIST 2006 MT Evaluationで1位) Aحدث النشرات

في خريف عام 2007، سيجتمع قادة الحكومات في مقر الأمم المتحدة في نيويورك للتوصل إلى حلول وتقديم توصيات لحماية الأطفال في الحرب. سينظر هؤلاء القادة فيما تم إحرازه من تقدم منذ تقرير عام 1996 الذي أعدته امرأة من الموزمبيق اسمها غراسا ماثيل. قدم التقرير توصيات مبتكرة حول ما يمكن - وما يجب - القيام به. وفي سعيه لإيصال أصوات الشباب، أجرى موقع أصوات الشباب مسحا للشباب يستطلع آراءهم حول ما يجب القيام به لاحترام حقوق الأطفال المتأثرين بالنزاع المسلح وحمايتهم وتلبيحتهم. وسيتم إدراج إجابات الشباب - من خلال المسح والمجموعات البورية - في التقرير الذي سيكون وثيقة رئيسية في النقاشات القادمة في الأمم المتحدة. وتقدم النشرة الإخبارية لهذا الشهر لمحة أولى حول رأي الشباب في أسباب الحرب وتبعاتها، وكيفية التأقلم معها، وتوصياتهم لقادة الحكومات.

↓

The latest releases we

In autumn 2007, will meet government leaders at United Nations Headquarters in New York to find solutions and make recommendations to protect children in war. Will these



参考文献 * = ACL Best Paper Award

Brown, P.F. et al. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 19(2):263-311.

*Chiang, D. 2007. Hierarchical phrase-based translation. *Computational Linguistics* 33(2):201-228.

*Germann, U. et al. 2001. Fast decoding and optimal decoding for machine translation. In Proc. of ACL 2001.

Koehn, P. et al. 2003. Statistical phrase-based translation. In Proc. of HLT-NAACL 2003.

Koehn, P. et al. 2007. Moses: open source toolkit for statistical machine translation. In Proc. of ACL 2007.

Marcu, D. and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In Proc. of EMNLP-2002.

Och, F.J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19-51.

Och, F.J. 2003. Minimum error rate training in statistical machine translation. In Proc. of ACL 2003.

*Och, F.J. and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In Proc. of ACL 2002.

Papineni, K. et al. 2002. Bleu: a method for automatic evaluation of machine translation. In Proc. of ACL 2002.

Weaver, W. 1947. The letter of W. Weaver appears in his paper "Translation" which was published in the book of "Machine Translation of Languages: Fourteen Essays" edited by W.N. Locke and A.D. Booth, 1955. (Also the paper is reprinted in "Readings in Machine Translation" edited by S Nirenburg et al., 2003.)

Wu, D. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics* 23(3):377-404.

Yamada, K., and K. Knight. 2001. A syntax-based statistical translation model. In Proc. of ACL 2001.

➡ **統計的機械翻訳の概要**

- 言語モデル: ngramモデル
- 翻訳モデル: フレーズ翻訳モデル
- デコーダ: Multistack & Beam-search

基本モデル: Noisy Channel Model

$$\hat{e} = \arg \max_e P(e | f)$$

$$= \arg \max_e \frac{P(f | e)P(e)}{P(f)}$$

$$= \arg \max_e P(f | e)P(e)$$

↓
↓
↓

デコーダ
(原言語に対して
意味的に等価で
流暢な目的言語を
探す)

翻訳モデル
(意味的等価性の
モデル)

言語モデル
((目的)言語らしさの
モデル)

$f = f_1^m = f_1, f_2, \dots, f_m$

$e = e_1^l = e_1, e_2, \dots, e_l$

確率論の復習

- P(a): 確率
 - aという事象が起きる確率
 - 例: P(曇り) = 1/3
- P(a, b): 同時確率
 - aとbという事象が同時に起きる確率
 - 例: P(今日曇り, 明日曇り) = 1/6
- P(a | b): 条件付確率
 - bという事象が起こった状態で, aという事象が起きる確率
 - 例: P(明日曇り | 今日曇り) = 1/2
 - 定理(定義): $P(a | b) = P(a, b) / P(b) = P(b | a) P(a) / P(b)$
- $P(a) = \sum_b P(a, b)$: 周辺確率 ベイズの定理

2つのモデル

$f = \text{He ate soup with a spoon.}$ 英日方向
↓ ↓
 $f e$

確からしさ ↑

ヘリウム スープ スプーン
彼 食べた と
は 飲んだ を で
が に
位置

e : 翻訳候補

- ヘリウム食べたスープとスプーン
- 彼は飲んだスープでスプーン
- ...
- ヘリウムはスプーンでスープを飲んだ
- ...
- 彼はスプーンでスープを飲んだ...

翻訳モデル $P(f|e)$ × 言語モデル $P(e)$

「同じ意味で かつ 日本語らしい」

デコーダ: 翻訳システム本体

$f = \text{He ate soup with a spoon.}$

デコーダ

$\hat{e} = \arg \max_e P(f|e)P(e)$

- $e_1 = \text{ヘリウム食べたスープとスプーン}$
- $e_2 = \text{彼は食べたスープでスプーン}$
- $e_3 = \text{ヘリウム飲んだスプーンとスープ}$
- ...
- $e_{28459} = \text{ヘリウムはスプーンでスープを飲んだ}$
- $e_{28460} = \text{彼はスプーンでスープを飲んだ}$
- $e_{28461} = \dots$

典型例 $20^{20} \times 20!$

$\hat{e} = \text{彼はスプーンでスープを飲んだ}$

翻訳候補数 = $M^{|f|} \times |f|!$

単語ごとの翻訳候補数 M f の文長

以降のスライド

- 午前
 - 統計的機械翻訳の概要
 - 言語モデル
 - 翻訳モデル
 - デコーダ
 全体概要
- 午後
 - フレーズ翻訳モデルの推定
 - 方向のある単語対応 (IBMモデル & GIZA++)
 - 対称化単語対応
 - フレーズペア抽出
 モデル推定
 - 評価指標 (BLEU)
 - Minimum Error Rate Training (MERT)
 その他

講義では省略

統計的機械翻訳の概要

- 言語モデル: ngramモデル
- 翻訳モデル: フレーズ翻訳モデル
- デコーダ: Multistack & Beam-search

ngramモデル

- ngramモデル = $n-1$ 重マルコフモデル

スライドに文字化け

$$P(e) = P(e_1, e_2, \dots, e_N)$$

$$= P(e_1)P(e_2 | e_1)P(e_3 | e_1, e_2) \dots P(e_N | e_1, e_2, \dots, e_{N-1})$$

$$= \prod_{i=1}^N P(e_i | e_1, e_2, \dots, e_{i-1})$$

$n-1$ 個前までの単語にのみ依存するとする

$$\approx \prod_{i=1}^N P(e_i | e_{i-n+1}, e_{i-n+2}, \dots, e_{i-1})$$

例: Bigram = 2gram: $P(e) \approx \prod_{i=1}^N P(e_i | e_{i-1})$

- $n=1$: Unigramモデル
- $n=2$: Bigramモデル
- $n=3$: Trigramモデル

$P(e = \text{"I think so"}) \approx P(I | w_0) \times P(\text{think} | I) \times P(\text{so} | \text{think})$

文頭単語 \leftarrow 7億回 \leftarrow Googleヒット数 \leftarrow 10万回

$P(e = \text{"My think so"}) \approx P(My | w_0) \times P(\text{think} | My) \times P(\text{so} | \text{think})$

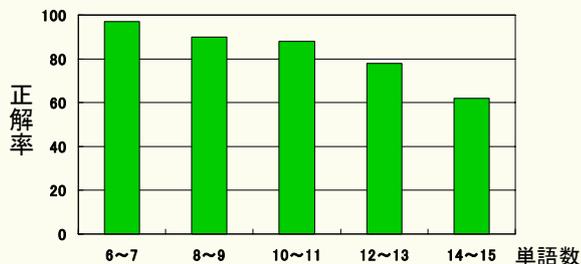
ngramモデルのパワー

- Bag translation
 - 文の単語(文字)を乱数ででたらめに入れ換える
 - ngramモデルで元に戻る
 - 最も高い確率を与える並びを出力とする
- 例:
 - 「彼はスプーンでスープを飲んだ。」
 - 「飲んだ, ., を, は, スープ, 彼, スプーン,」
 - 「彼はスプーンでスープを飲んだ。」??

Bag translationの実験結果 1/2

• 使用モデル

- 単語Trigram (Back-off Smoothing)
- 訓練データ: 毎日新聞5年分 ← 最尤推定の改良



19

Bag translationの実験結果 2/2

入力例:

を、て、その、、た、うかがつ、。、い、絶対、で、チャンス、斉藤、は、の、自信

- 完全に正しい復元 (14~15単語の文で60%強)
 - 元の文: 斉藤は絶対の自信で、そのチャンスをうかがっていた。
 - 復元文: 同上
- おいしい!
 - 元の文: その後結婚した妻や2人の子どもも失った。
 - 復元文: その後結婚した2人の妻や子どもも失った。
- 誤った復元
 - 元の文: 更新には警視庁の担当者が皇居に出張してくる。
 - 復元文: 警視庁の担当者には更新が皇居に出張してくる。

20

言語モデルの研究は終わっている?

• ngramモデルの推定

- 問題: データのスパースネス
 - trigramモデル&語彙10万単語 → 100,000³=1[Peta]
- 最尤推定の改良
 - 工学的な手法としてはほぼ確立している
 - back-off smoothing や Interpolation

• 研究

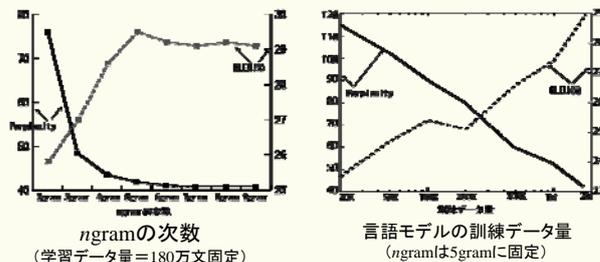
- 音声認識の言語モデル
 - 約10年ほど前に研究は下火
 - 言語モデルの性能改善が音声認識率の改善に結びつかない
- SMTの場合
 - 言語モデルの性能改善は音声認識よりもはるかに翻訳性能の改善に効果がある(?)

21

言語モデルの性能は重要

- 言語モデルの性能 = Perplexity (小さいほどよい)
- 翻訳システムの性能 = BLEU (大きいほどよい)

英日特許翻訳, フレーズ翻訳モデル=180万文ペア (NTCIR7), Reorderingモデルなし



ngramの次数

(学習データ量=180万文固定)

言語モデルの訓練データ量

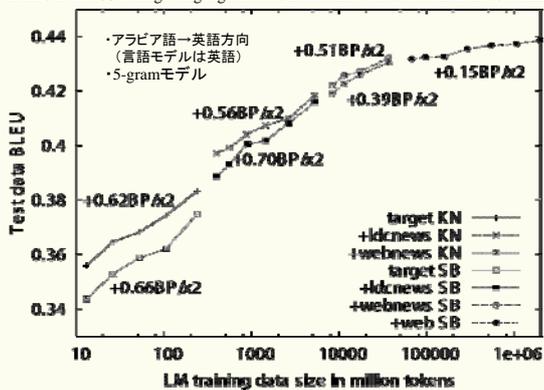
(ngramは5gramに固定)

PerplexityとBLEUはきれいに反比例する

22

Google ngram

T.Brants et al. 2007. Large language models in machine translation. EMNLP-CoNLL2007. より



ベースシステム: NIST 2006 Machine Translation Evaluationで1位だったGoogleのシステム

統計的機械翻訳の概要

- 言語モデル: ngramモデル
- 翻訳モデル: フレーズ翻訳モデル
- デコーダ: Multistack & Beam-search

P.Koehn, F.Och and D. Marcu. 2003. Statistical phrase-based translation, NAACL-2003, pp.48-54.
D. Marcu and W.Wong. 2002. A phrase-based, joint probability model for SMT, EMNLP-2002, pp.133-139.

をベースに厳密かつ分かりやすく(?) 改善

- Phrase Alignment (アラインメント or 対応)
- フレーズ翻訳モデル
- 歪モデル

24

小さい単位への分解

- $P(f|e)$: e が f に翻訳される確率

$f = \text{the remains are the same as in fig. 12.}$
 $e = \text{それ以外は 図12と同様である。}$

↓

$P(f|e)??$

- 文や単語のように順序関係がない離散確率変数に対する確率分布は表で与えるしか方法がない
- 文(の組)は、表にするには多すぎる

↓

「単語または単語列の対訳対の翻訳確率」を組み合わせる

↓

「フレーズ」と呼ぶ

25

Alignment 1/3

- フレーズの翻訳確率 $t(\bar{f}|\bar{e})$ が与えられていて、
 $t(\text{the remains} | \text{それ以外は}) = 0.08, t(\text{the same as in} | \text{と同様}) = 0.3, \dots$
- かつ、あるフレーズ対応 a (Alignment) が与えられれば、

$f = \text{the remains are the same as in fig. 12}$
 $a \rightarrow$ |
 $e = \text{それ以外は 図12と同様である}$

↓

$P(f|e,a) \doteq P(\text{the remains} | \text{それ以外は})P(\text{are} | \text{である})$
 $P(\text{the same as in} | \text{と同様})P(\text{fig. 12} | \text{図12})$

26

Alignment 2/3

- Alignment a の表現

$f = \bar{f}_1^{m=4} = \text{the remains are the same as in fig. 12}$
 $e = \bar{e}_1^{m=4} = \text{それ以外は 図12と同様である}$

それ以外は 図12と同様である

$a = a_1^m = a_1, a_2, \dots, a_m$ と表現し、 a_j は $f_j \leftrightarrow e_{a_j}$ を意味する。

上記の例: $a = a_1^4 : a_1=1, a_2=4, a_3=3, a_4=2$

$\bar{f}_2 \leftrightarrow \bar{e}_{a_2=4}$

前ページは $P(f|e,a = a_1^m) \approx \prod_{i=1}^m P(\bar{f}_i | \bar{e}_{a_i})$ と書ける

27

Alignment 3/3

- $P(f|e,a)$ から $P(f|e)$ はどうやって求める? reorderingモデル

$$P(f|e) = \sum_a P(f,a|e) = \sum_a P(f|e,a)P(a|e)$$

$$P(f|e,a = a_1^m) \approx \prod_{i=1}^m P(\bar{f}_i | \bar{e}_{a_i})$$

$f = \text{the remains are the same as in fig. 12}$
 a_1 |
 $e = \text{それ以外は 図12と同様である}$

$P(f, a_1|e)$

$f = \text{the remains are the same as in fig. 12}$
 a_2 |
 $e = \text{それ以外は 図12と同様である}$

$P(f, a_2|e)$

$f = \text{the remains are the same as in fig. 12}$
 a_3 |
 $e = \text{それ以外は 図12と同様である}$

$P(f, a_3|e)$

合計が $P(f|e)$

28

MosesのReorderingモデル

$$P(f|e) = \sum_a P(f,a|e) = \sum_a P(f|e,a)P(a|e)$$

$f = \text{the remains are the same as in fig. 12}$
 $e = \text{それ以外は 図12と同様である}$
 $a_1=1 \quad a_4=2 \quad a_3=3 \quad a_2=4$
 $(m) \quad (d) \quad (s) \quad (s)$
 $(a_0=0 \text{ と考える})$

$P(a|e) = \prod_{i=1}^m P(a_i | e, a_{i-1})$
 $\approx \prod_{i=1}^m P(a_i | \bar{e}_{a_i}, a_{i-1})$
 $\approx \prod_{i=1}^m P(\text{orientation} | \bar{e}_{a_i})$

$$\text{orientation} = \begin{cases} \text{monotone}(m), & \text{if } // \\ \text{swap}(s), & \text{if } \times \\ \text{discontinuous}(d), & \text{otherwise.} \end{cases}$$

実習のときのオプション = msd-bidirectional-fe

lexicalised msd モデル

C.Tillman. 2004. A Unigram Orientation Model for Statistical Machine Translation. HLT-NAACL2004.
 29 P.Koehn et al. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. IWSLT2005.

統計的機械翻訳の概要

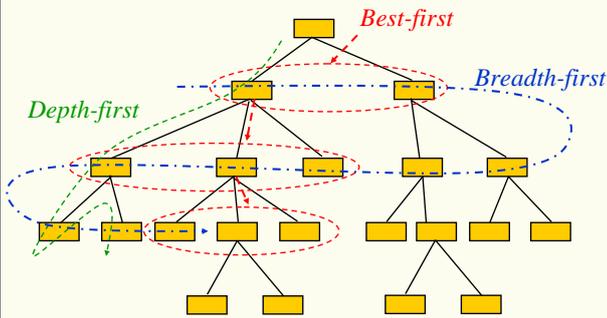
- 言語モデル: ngramモデル
- 翻訳モデル: フレーズ翻訳モデル
- デコーダ: Multistack & Beam-search

Pharaoh, Moses
 (フレーズベースの代表的decoder)

30

基本探索手法

- 知識なし探索手法: *Depth-first*, *Breadth-first*
- 知識あり探索手法: *Best-first*

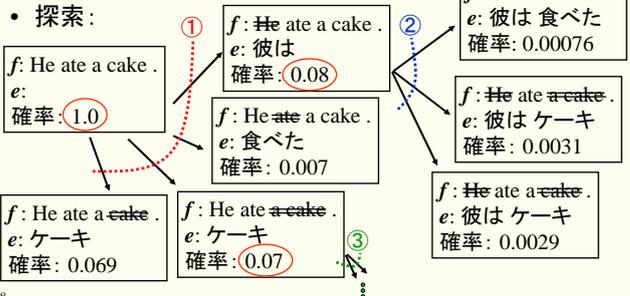


37

Best-first search

- 仮説: f : すでに翻訳した部分
 e : 翻訳候補 (先頭から途中まで)
確率: 翻訳した部分の確率 $f(h)$

Stack decoderと
呼ばれることもある



38

Future cost

- $f(h)$ の問題
 - $f(h)$ は部分翻訳確率の*積*なので、翻訳が進むほど $f(h)$ が小さくなる
 - 全体で見たとき、翻訳が進んだ仮説は展開されない傾向にある
 - Breadth-first search*になってしまう

評価値を翻訳された部分の長さに依存しないようにする
 ⇒ 未翻訳部分の翻訳コスト(確率)を推測して加える(積)
 $Future\ cost = g(h)$

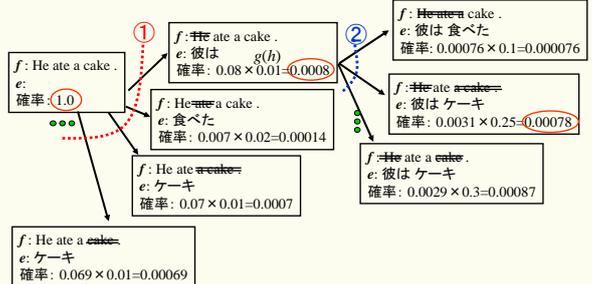
$g(h)$ がある条件を満たせば A^* search になる

39

Best-first + Future cost

- Future cost*の例 (Moses)

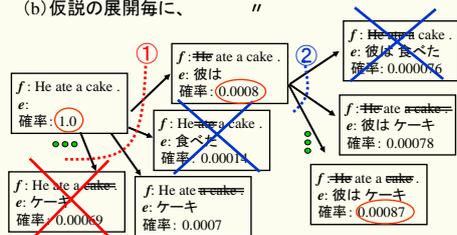
- 未翻訳部分のフレーズ翻訳確率
 - 言語モデルはフレーズ内部のみ
 - reorderingモデルは無視
- 動的計画法 (DP or Viterbi) で
高速計算できる
- 未翻訳部分の確率を
最大とする組合せ



40

Beam-search

- Best-first search*の問題
 - すべての展開仮説を保持するのは困難
 - 枝刈り
 - (a) 全体で、上位 b 個 (*beam*幅と呼ばれる)しか仮説を保持しない
 - (b) 仮説の展開毎に、



例えば、全体で上位3個しか保持しない場合 (a).

41

Multistack & Beam-search

- 枝刈りつき *Best-first search* の問題
 - Future cost* が真の *cost* より小さめの場合
 - 翻訳が進んでいない仮説を切りすぎる
 - Future cost* が真の *cost* より大きめの場合:
 - 翻訳が進んだ仮説を切りすぎる

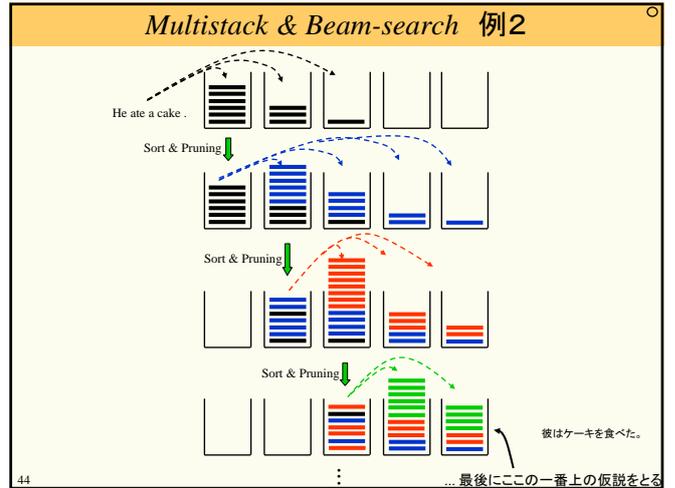
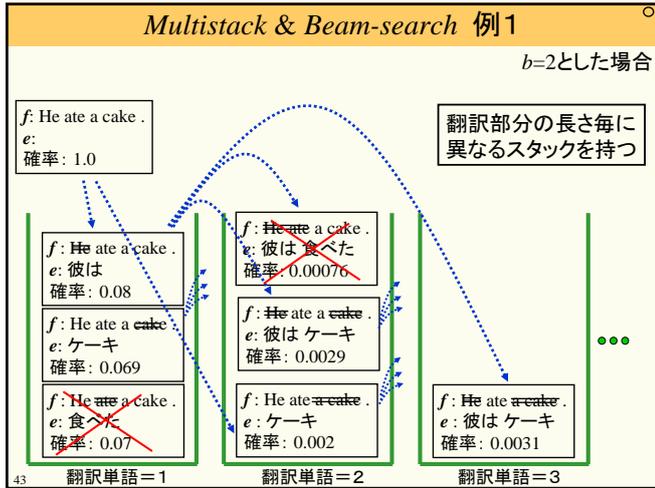
結局、未翻訳部分の長さが異なる仮説の優劣を比較すること自体に無理がある

未翻訳部分の長さごとに分けて枝刈りを行う

Multistack & Beam-search

現在の主流

42



フレーズ翻訳モデルの推定

➡ フレーズ翻訳モデル推定の手順

- 方向のある単語対応
- 対称化単語対応
- フレーズペア抽出

45

フレーズ翻訳モデル推定の手順

- 手順
 - 方向のある単語対応
 - GIZA++ (Och&Ney 2003)
 - $P(f|e)$ に対してIBMモデルを推定しながら、対訳データの各ペアに対して、IBMモデルによる最尤の単語alignment ($\hat{a} = \arg \max_a P(f, a|e)$)を計算する。
 - 逆方向 $P(e|f)$ に対しても同じことをする
 - 対称化単語対応: Symmetrized word alignment
 - ヒューリスティクス
 - 各ペアに対して、両方向の単語alignmentを用いて対称単語対応を得る
 - 例えば: Grow-diag-final-and (Koehn et al. 2005)
 - フレーズ抽出
 - 一貫した対訳フレーズ $\langle \bar{f}, \bar{e} \rangle$ を抽出する
 - フレーズ翻訳確率 $t(\bar{f}|\bar{e})$ を計算する

46

フレーズ翻訳モデル推定の手順: 例

(1) 方向のある単語対応: $P(f=日|e=英)$
それ以外は図12と同様である
 e_0 the remains are the same as in fig. 12

(1) 方向のある単語対応: $P(f=英|e=日)$
それ以外は図12と同様である
the remains are the same as in fig. 12

(2) 対称化単語対応
それ以外は図12と同様である
the remains are the same as in fig. 12
積: \ のみ
和: \ と
その他

(3) フレーズ抽出
それ以外は図12と同様である
the remains are the same as in fig. 12

47

フレーズ翻訳モデルの推定

- フレーズ翻訳モデル推定の手順
- ➡ 方向のある単語対応 (IBMモデル&GIZA++)
- 対称化単語対応
- フレーズペア抽出

48

翻訳モデル: $P(f|e)$

• IBMモデル (単語に基づく翻訳モデル)

- P.F. Brown et. al. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), pages 263-311.
(<http://acl.ldc.upenn.edu/J/J93/J93-2003.pdf>)

- Model 1 ~ 5

- $P(f|e)$ の近似方法の違い
- だんだん精巧なモデルとなる

- パラメータ推定

- データ: 文対応の対訳コーパス
- 推定手法: 最尤推定 (EMアルゴリズム)

49

Alignment 1/4

- $P(f|e)$: e が f に翻訳される確率

$f = \text{CE NE EST PAS CLAIR}$
 $e = \text{It is not clear}$



$P(f|e) ??$

- 文や単語のように順序関係がない離散確率変数に対する確率分布は表で与えるしか方法がない
- 文(の組)は、表にするには多すぎる



「単語対の翻訳確率」を組み合わせる

50

Alignment 2/4

- 単語の翻訳確率 $t(f|e)$ が与えられていて、
 $t(\text{CE}|\text{it}) = 0.8, t(\text{NE}|\text{it}) = 0.00001, \dots$ (f と e は単語)

- かつ、ある単語対応 a (Alignment) が与えられれば、

$f = \text{CE NE EST PAS CLAIR}$
 $e = \text{It is not clear}$



$$P(f|e, a) \doteq P(\text{CE}|\text{it})P(\text{NE}|\text{not})P(\text{EST}|\text{is})P(\text{PAS}|\text{not})P(\text{CLAIR}|\text{clear})$$

- $P(f|e, a)$ から $P(f|e)$ はどうやって求める?

$$P(f|e) = \sum_a P(f, a|e) = \sum_a P(f|e, a)P(a|e)$$

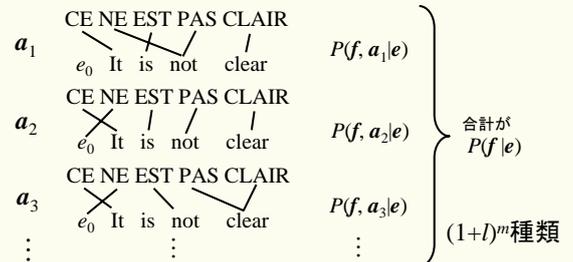
これを一様分布とおくのがModel1

51

Alignment 3/4

- IBM model におけるAlignmentの条件

- f 側の各単語は e 側に対応する単語をそれぞれ1つもつ
- e 側の各単語は f 側の $0 \sim m$ 単語に対応する可能性がある
- e 側には空単語 e_0 が存在すると仮定する
- f 側のある単語に対応する e 側単語がない場合は、 e_0 に対応させる



52

Alignment 4/4

- Alignment a の表現

$f = f_1^m = \begin{matrix} \text{CE} & \text{NE} & \text{EST} & \text{PAS} & \text{CLAIR} \\ f_1 & f_2 & f_3 & f_4 & f_5 \end{matrix}$

$e = e_0^l = \begin{matrix} e_0 & e_1 & e_2 & e_3 & e_4 \\ e_0 & \text{It} & \text{is} & \text{not} & \text{clear} \end{matrix}$

$a = a_1^m = a_1, a_2, \dots, a_m$ と表現する。

a_j は $f_j \leftrightarrow e_{a_j}$ を意味する。

上記の例: $a = a_1^5 : a_1=1, a_2=3, a_3=2, a_4=3, a_5=4$

$f_2 \leftrightarrow e_{a_2=3}$

53

IBM Model 1

- IBM Model 1 は、 $t(f|e)$ だけを用いた翻訳モデル

- ◆ 特長: EMアルゴリズムで大域的に最適なモデル推定が可能 (最尤推定)

$$P(f|e) = \sum_a P(f, a|e)$$

Model 1 では一様分布とする

$$= \sum_a P(a|e)P(f|e, a)$$

$$\approx \sum_a \varepsilon \prod_{j=1}^m t(f_j|e_{a_j})$$

$$= \varepsilon \sum_{a_1=0}^l \sum_{a_2=0}^l \dots \sum_{a_m=0}^l \prod_{j=1}^m t(f_j|e_{a_j})$$

54

翻訳モデルの推定: 最尤推定

- 対訳コーパス: $D =$ 対訳ペア $\langle f, e \rangle$ の集合
- 尤度 $= \prod_{\langle f, e \rangle} P(f | e)$
 - 各対訳ペアが独立に生じたと仮定する
- 最尤推定

尤度を最大とするパラメータを求める

対数尤度 $L(D | \theta) = \log \prod_{\langle f, e \rangle} P(f | e)$

$$= \sum_{\langle f, e \rangle} \log P(f | e)$$

ただし、全対訳ペアで共起しない単語ペアを確率ゼロと仮定すると計算する必要はない。この場合パラメータは数百万。

$$\hat{\theta} = \arg \max_{\theta} L(D | \theta)$$

Model1の場合 $t(f|e)$ 最低でも10万×10万=100億のパラメータ

IBM Model 1の推定

- 対数尤度

$$L(D | \theta = \{t(f | e)\}) = \sum_{\langle f, e \rangle} \log P(f | e)$$

$$= \sum_{\langle f, e \rangle} \log \varepsilon \prod_{a_1=0}^l \sum_{a_m=0}^m \prod_{j=1}^m t(f_j | e_{a_j})$$

$$= const. + \sum_{\langle f, e \rangle} \log \sum_{a_1=0}^l \sum_{a_m=0}^m \prod_{j=1}^m t(f_j | e_{a_j})$$

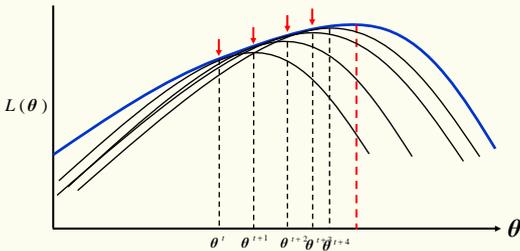
目的関数(最大化) $L(\theta) = \sum_{\langle f, e \rangle} \log \sum_{a_1=0}^l \sum_{a_m=0}^m \prod_{j=1}^m t(f_j | e_{a_j})$

制約条件 $\sum_f t(f | e) = 1$ for $\forall e$.

IBM Model 1の推定

T. Minka, 1998. Expectation-Maximization as lower bound maximization. <http://research.microsoft.com/~minka/papers/em.html>

- EMアルゴリズム
 - 現在の θ^{old} を $L(\theta^{old}) \leq L(\theta^{new})$ を満たすような θ^{new} に更新
 - ポイント: lower boundの最大化を繰り返す



EMアルゴリズムは $\sum_k p_k r_k \geq \prod_k r_k^{p_k}$ を使う ($\sum_k p_k = 1$)

(幾何平均は算術平均より小さいか等しい)

IBM Model 1の推定

$$L(\theta^{new}) = \sum_{\langle f, e \rangle} \log \sum_{a_1=0}^l \sum_{a_m=0}^m \prod_{j=1}^m t^{new}(f_j | e_{a_j})$$

$$= \sum_{\langle f, e \rangle} \log \prod_{j=1}^m t^{new}(f_j | e_i)$$

当たり前? (説明は次ページ)

$$= \sum_{\langle f, e \rangle} \sum_{j=1}^m \log \sum_{i=0}^l t^{new}(f_j | e_i)$$

$$= \sum_{\langle f, e \rangle} \sum_{j=1}^m \log \sum_{i=0}^l p_{ji}(\theta^{old}) \frac{t^{new}(f_j | e_i)}{p_{ji}(\theta^{old})}$$

$$\geq \sum_{\langle f, e \rangle} \sum_{j=1}^m \log \prod_{i=0}^l \left(\frac{t^{new}(f_j | e_i)}{p_{ji}(\theta^{old})} \right)^{p_{ji}(\theta^{old})}$$

$$= \sum_{\langle f, e \rangle} \sum_{j=1}^m \sum_{i=0}^l p_{ji}(\theta^{old}) \log t^{new}(f_j | e_i) + const.$$

$$p_{ji}(\theta^{old}) = \frac{t^{old}(f_j | e_i)}{\sum_{k=0}^l t^{old}(f_j | e_k)}$$

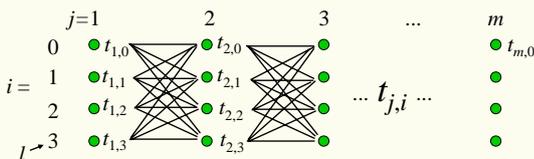
$$= L'(\theta^{new})$$

Lower bound, これを最大化する

参考: $\sum_{a_1=0}^l \sum_{a_m=0}^m \prod_{j=1}^m t(f_j | e_{a_j}) = \prod_{i=1}^m \sum_{j=1}^l t(f_j | e_i)$

$$\sum_{a_1=0}^l \sum_{a_m=0}^m \prod_{j=1}^m t(f_j | e_{a_j}) = \prod_{a_1=0}^l \sum_{j=1}^m t_{j,a_1}$$

- 上式は、以下の図で「パス ($j=1 \sim m$) 上の $t_{j,i}$ の積」をあらゆるパスについて和をとったものである。



- 同じことは以下の式でもできる。

$$(t_{1,0} + t_{1,1} + t_{1,2} + t_{1,3}) \times (t_{2,0} + t_{2,1} + t_{2,2} + t_{2,3}) \times \dots$$

$$= \prod_{j=1}^m \sum_{i=0}^l t_{j,i} = \prod_{i=1}^m \sum_{j=1}^l t(f_j | e_i)$$

IBM Model 1の推定

- 等式制約条件: $\sum_f t(f | e) = 1$ for $\forall e$.

- ラグランジュ関数

$$h(\theta^{new}) = L'(\theta^{new}) + \sum_e \lambda_e (\sum_f P(f | e) - 1)$$

$$= \sum_{\langle f, e \rangle} \sum_{j=1}^m p_{ji}(\theta^{old}) \log t^{new}(f_j | e_i) + \sum_e \lambda_e (\sum_f P(f | e) - 1)$$

- 最大化

$$\frac{\partial h(\theta^{new})}{\partial t(f | e)} = 0$$

$$t^{new}(f | e) \propto \sum_{\langle f, e \rangle} \frac{t^{old}(f | e)}{t^{old}(f | e_0) + L + t^{old}(f | e_i)} \sum_{j=1}^m \delta(f, f_j) \sum_{i=0}^l \delta(e, e_i)$$

クロネッカーの δ ($f=f_j$ のとき1, 他は0)

目的関数のもとと上に凸なので、最適解が求まる。

IBM Model 1 推定の解釈

- Alignmentが与えられていたら？

$$t(\text{CE}|\text{It}) = \frac{\text{It} \leftrightarrow \text{CEの出現回数}}{\text{It} \leftrightarrow *の出現回数}$$

Alignmentの数(例では5)だけ投票
- Alignmentが与えられていない場合

$$t(\text{CE}|e_0), t(\text{CE}|\text{It}), t(\text{CE}|\text{is}), t(\text{CE}|\text{not}), t(\text{CE}|\text{clear})$$

に分配して投票

$$t^{new}(f|e) \propto \sum_{\langle f, e \rangle} \frac{t^{old}(f|e)}{t^{old}(f|e_0) + L + t^{old}(f|e_1) + \dots} \sum_{j=1}^m \delta(f, f_j) \sum_{i=0}^l \delta(e, e_i)$$

61

IBM Model 1~5

- IBM Model 1~5
 - Model 1: translation prob. $t(f|e)$
 - Model 2: Model 1 + permutation prob. (絶対位置)
 - Model 3: Model 2 + fertility prob. $n(\phi|e)$
 - Model 4: Model 3 の permutation prob. を改良 (相対位置)
 - Model 5: Model 4 の permutation prob. をさらに改良
- Model 4: $P(f, a|e) = \text{fertility} \times \text{translation} \times \text{permutation}$
 - 単語対応数 単語対応 交換
 - fertility probability
 - 目的言語のある単語 e_i が原言語の何単語 (ϕ_i) に対応するか
 - translation probability
 - 目的言語のある単語 e_i が原言語の単語 f_{ik} に対応する(訳される)確率
 - permutation probability
 - 目的言語のある単語 e_i に対応する原言語の単語 f_{ik} の原言語文中での位置 j の確率
 - (直前の目的言語単語 e_{i-1} に対応する原言語単語からの相対位置: π_{ik})

62

IBM Model 4

- $P(f, a|e) = \text{fertility} \times \text{translation} \times \text{permutation}$
 - 単語対応数 単語対応 交換

目的言語単語 e_i : $e_0 \ e_1 \ e_2 \ e_3 \ e_4$

fertility ϕ_i : 0 1 1 2 1 $\rightarrow n(\phi_i|e_i)$

translation f_{ik} : - ce est ne.pas clair $\rightarrow t(f_{ik}|e_i)$

permutation π_{ik} : - 1 2 -1,2 2 $\rightarrow d(\pi_{ik}|c(e_{i-1}, f_{ik}))$

notの場合の例

63

推定実験 1/2

- 実験条件
 - コーパス: Hansards corpus 1,778,620英仏文ペア
 - 語彙: 英 42,005単語 + 1, 仏 58,016単語
 - モデルパラメータ: (Model 5)
 - $t(f|e)$: $42,006 \times 58,016 = 2,437,020,096$
文ペア内に共起した単語だけを考慮 $\rightarrow 25,427,016$
 - $n(\phi|e)$: $5(?) \times 42,006$
 - $d(\pi|\dots)$:

P.F. Brown et. al. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), pages 263-311.

64

推定実験 2/2

not

$t(\text{ne}|\text{not})=0.497$ $n(2|\text{not})=0.735$
 $t(\text{pas}|\text{not})=0.442$ $n(0|\text{not})=0.154$
 $t(\text{non}|\text{not})=0.029$ $n(1|\text{not})=0.107$
 ⋮

oil

$t(\text{petrole}|\text{oil})=0.442$
 $t(\text{petrolieries}|\text{oil})=0.138$
 $t(\text{petroliere}|\text{oil})=0.109$
 $t(\text{le}|\text{oil})=0.054$
 ⋮

the

$t(\text{le}|\text{the})=0.497$ $n(1|\text{the})=0.746$
 $t(\text{la}|\text{the})=0.207$ $n(0|\text{the})=0.254$
 $t(\text{les}|\text{the})=0.155$
 ⋮

P.F. Brown et. al. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), pages 263-311.

65

フレーズ翻訳モデルの最尤推定

単語翻訳モデル (IBM Model 1)

$$P(f|e) = \sum_a P(f, a|e)$$

$$= \sum_a P(a|e) P(f|e, a)$$

$$\approx \varepsilon \sum_a \prod_{i=1}^m t(f_i|e_{a_i}) (l+1)^m$$

$$= \varepsilon \prod_{i=1}^m \sum_{j=0}^l t(f_i|e_j) (l+1)^m$$

↓

- 現実な計算量かつ大域的最適化を行える美しいアルゴリズム

単語モデルと最尤推定は相性がよい

フレーズ翻訳モデル

$$P(f|e) = \sum_a P(f, a|e)$$

$$= \sum_a P(a|e) P(f|e, a)$$

$$\approx \varepsilon \sum_a \prod_{i=1}^m P(\bar{f}_i|\bar{e}_{a_i})$$

↓

- 計算量を下げられない
- 近似が必要となる
- Alignmentの可能性が単語のときより膨大
- フレーズの階層性の問題

↓

ヒューリスティックス

66

最尤単語アラインメント: GIZA++

- GIZA++
 - IBM Model (+ HMM) モデルの推定
 - 訓練用対訳文ペアに対して最尤単語アラインメント
Viterbiアラインメント

$$\hat{a} = \arg \max_a P(f, a | e)$$

最尤単語アラインメントの計算方法

- シンプルなモデルのViterbiアラインメント(厳密解)を求める
 - f 側単語に同期させてViterbiアルゴリズムを適用
 - Model1, Model2, HMM
- 上のViterbiアラインメントを初期値として局所探索でよりよいアラインメントを探す
(Fertilityが難点)
- Model3以上

Och, F.J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19-51.

67

フレーズ翻訳モデルの推定

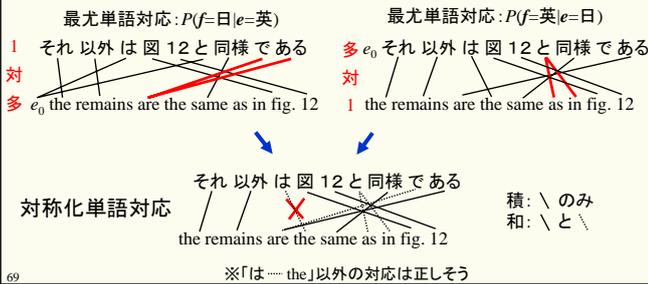
- フレーズ翻訳モデル推定の手順
- 方向のある単語対応
- 対称化単語対応
- フレーズペア抽出

68

対称化単語対応

目的

- (1) IBMモデルは方向のある1対多のalignment
→ 両方向の1対多のalignmentにしたい
- (2) IBMモデルのalignmentにも誤りが多い
→ 正しい対応だけを残したい



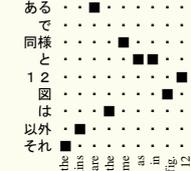
69

対称化単語対応

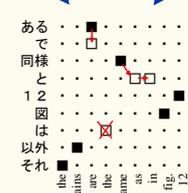
最尤単語対応: $P(f=日|e=英)$



最尤単語対応: $P(f=英|e=日)$



対称化単語対応



積: ■のみ
和: ■と□
grow-diag: Xを除く■と□

P. Koehn et al. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. IWSLT2005.

70

対称化単語対応: ヒューリスティックス

- 基本3種
 - 積 (intersection): 両方向共にある対応点のみを用いる
alignment point
 - 高precision, 低recall
 - 和 (union): 両方向の対応点をすべて使う
 - 低precision, 高recall
 - 成長 (grow): 積と和の中間
 - 積からスタート (和の対応点)
 - すでに採用した対応点の周りに候補の対応点を加えていく
 - 周り: grow → 縦・横, grow-diag → 縦・横・対角
- 最終処理(追加)3種 (和の対応点)
 - 対応がない単語に関する候補対応点を加える
 - final: 少なくとも片方の言語の単語の単語対応がない場合
 - final-and: 両側言語の単語に単語対応がない場合
 - 最終処理を行わない

71

フレーズ翻訳モデルの推定

- フレーズ翻訳モデル推定の手順
- 方向のある単語対応
- 対称化単語対応
- フレーズペア抽出

72

一貫したフレーズペア

対称化単語対応 (grow-diag) 定義:
 フレーズ=連続する単語列
 フレーズペア=
 英日のフレーズの対応

ある
 で
 同様
 と
 1 2
 図
 は
 以外
 それ

the remains are the same as in fig. 12

左の図のalignmentポイントを
 囲む四角で任意のフレーズペア
 を表現できる。
 例: 左の□は「と同様」と
 “the same as in”のペア

一貫したフレーズペア=
 灰色部分にAlignment
 ポイントがないフレーズ

一貫したフレーズペアをすべて取り出す

一貫しないフレーズペア

対称化単語対応 (grow-diag) 対称化単語対応 (grow-diag)

ある
 で
 同様
 と
 1 2
 図
 は
 以外
 それ

the remains are the same as in fig. 12

一貫したフレーズペア=灰色部分にAlignment
 ポイントがないフレーズ

一貫したフレーズペア (別の図)

対称化単語対応 (grow-diag) 対称化単語対応 (grow-diag)

それ以外は図12と同様である
 the remains are the same as in fig. 12

と同様
 the same as in
 一貫している

それ以外は図12と同様である
 the remains are the same as in fig. 12

と同様
 the same as in
 一貫しない

一貫したフレーズペア=
 フレーズペア内の単語対応から外に出て行く対応がない

フレーズペア抽出

対称化単語対応 (grow-diag) 抽出されるフレーズペア

ある
 で
 同様
 と
 1 2
 図
 は
 以外
 それ

the remains are the same as in fig. 12

(それ, the) (以外, remains)
 (図, fig.) (12, 12) (同様, same)
 (と, as in) (同様, the same)
 (である, are) (は, fig.)
 (以外, are, remains)
 (図12, fig. 12) (である, are the)
 (それ以外, the remains)
 (と同様, same as in)
 (それ以外, the remains)
 (は図12, fig. 12)
 (図12と, as in fig. 12)
 (と同様, the same as in)
 (同様である, are the same)
 (は図12と, as in fig. 12)
 (と同様である, are the same as in)
 (図12と同様, same as in fig. 12)
 (図12と同様, the same as in fig. 12)
 (は図12と同様, same as in fig. 12)
 (は図12と同様, the same as in fig. 12)

翻訳性能評価

- 人手評価
- 自動評価
 - BLEU, WER, NIST
- 人手評価と自動評価の相関

評価方法の種類

- 人手評価 (manual evaluation)
 - 翻訳結果は最終的に人が読むことが多い
 - 「翻訳性能」を測る場合は人手評価が基本
 - 欠点: 高コスト(労力&時間)
- 自動評価 (intrinsic evaluation)
 - 人手をかけない評価
 - ・主に、正解翻訳例との一致率(様々)で定義されることが多い
 - 問題: 人手評価と正確に同じではない
 - 利点: 正解翻訳例さえあればあとは低労力・少時間
- 自動評価 (extrinsic evaluation)
 - 人手をかけない評価
 - ・翻訳システムを利用したより大きなタスクの達成率で評価
 - 例: NTCIR-7のPatent Translation TaskにおけるCLIR評価

人手評価

- Adequacy & Fluency (代表的)
 - Adequacy: 翻訳結果は正しく意味を伝えているか?
 - Fluency: 翻訳結果は言葉として自然か?
 - 複数の人間によるそれぞれ5段階評価の平均
- 例 (HLT-NAACL2006のSMT Workshop)

Adequacy	Fluency
5: All meaning	5: Flawless English
4: Most meaning	4: Good English
3: Much meaning	3: Non-native English
2: Little meaning	2: Disfluent English
1: None	1: Incomprehensible

79

自動評価 (intrinsic evaluation)

- BLEU: BiLingual Evaluation Understudy
 - 翻訳結果とreferenceとの1~4gramの一致率の幾何平均
 - 短い文にはペナルティ
 - 0.0~1.0で、大きいほどよい
- WER: Word Error Rate
 - 翻訳結果とreferenceの編集距離を正規化
 - 置換・挿入・削除誤りの割合
 - 音声認識におけるいわゆる「認識率」
 - 0.0~1.0で、小さいほどよい
- NIST metric: NIST (米国標準技術局) のBLEU
 - BLEUの幾何平均を重み付き幾何平均にする
 - 語順よりも単語訳の正しさに(特に内容語)を重視
 - 0.0~無限大で、大きいほどよい

$$BLEU = BP \cdot \sqrt[n]{\prod_{n=1}^4 p_n}$$

ペナルティ n ngramの一致率

PER=語順を無視したWER (Position-independent word Error Rate)

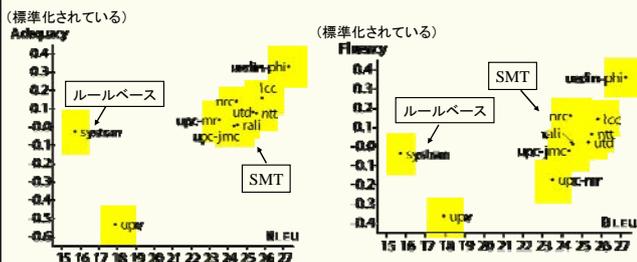
80

NIST: National Institute of Standards & Technology

人手評価とBLEUの相関

P.Koehn and C.Monz. 2006. Manual and automatic evaluation of machine translation between European languages. Proc. of the workshop on SMT, pages 102-121.

ドイツ語→英語



SMTシステム同士の比較では高い相関
異種システム(SMT以外?)の比較には使わない方がよい

81

参考文献など

- BLEU
 - オリジナルが以下
 - K.Papineni et al. 2002. *Blue: a method for automatic evaluation of machine translation*. Proc. of ACL2002, pages 311-318.
- WER
 - Wikipedia (英語版) の ``word error rate`` の項
 - 最初に定義(使用)したオリジナルな論文は不明
 - 比較的正確に書いてある論文は例えば以下
 - McCowan et al. 2005. On the use of information retrieval measures for speech recognition evaluation. IDIAP Research Report, IDIAP-RR 04-73.
- NIST metric
 - Wikipedia (英語版) の ``NIST (metric)`` の項
 - オリジナルは以下らしいがインターネット上にない(?) ので見たことがない
 - G.Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. HLT2002. (ACL AnthologyからHLT2002だけ全部欠落している)

82

チューニング

Minimum Error Rate Training

83

MERT: Minimum Error Rate Training

- log-linearモデル

$$\hat{e} \approx \arg \max_{e,a} P^l(e) P^l(f|e,a) P^l(a|e)$$

$$\hat{e} \approx \arg \max_{e,a} P^{\lambda_{LM}}(e) P^{\lambda_{FM}}(f|e,a) P^{\lambda_{RM}}(a|e)$$

$$\hat{e} \approx \arg \max_{e,a} \prod_k f_k(f,e,a)^{\lambda_k}$$

$$\hat{e} \approx \arg \max_{e,a} \log \prod_k f_k(f,e,a)^{\lambda_k}$$

$$= \arg \max_{e,a} \sum_k \lambda_k \log f_k(f,e,a)$$

本当は、基本モデルが、*Noisy Channel Model*から
↓
*Discriminative Model*へとドラスティックに変化しているが、本講義ではそこには立ち入らない。
詳しくは[Och&Ney2002]など

Minimum Error Rate Training (MERT): \hat{e} がよくなるように λ_k を決める

84

MERT : Minimum Error Rate Training

Minimum Error Rate Training (MERT): \hat{e} がよくなるように λ_k を決める

\hat{e} のよさ = 翻訳性能評価関数: BLEU, NIST, WER, ...
 いずれも正解例 e_{ref} との一致率
 $eval(\hat{e}, e_{ref})$ とおく(大きいほどよい)

MERT:

$$\hat{\lambda} = \lambda_k^n = \arg \max_{\lambda} eval(smt(\lambda), e_{ref})$$

$$smt(\lambda) = \hat{e} = \arg \max_e \sum_k \lambda_k \log f_k(e, a)$$

maxが2段階適用されている
 ので最適化は簡単ではない(→近似: 次ページ)

※BLEUは文の集合に対して定義されるので、 \hat{e} や e_{ref} は文集合となる

MERT : Minimum Error Rate Training

$$\hat{\lambda} = \lambda_k^n = \arg \max_{\lambda} eval(smt(\lambda), e_{ref})$$

$$smt(\lambda) = \hat{e} = \arg \max_e \sum_k \lambda_k \log f_k(e, a)$$

具体的な手続き

- (0) λ を初期値に設定する
- (1) 現在の λ でデコーダを走らせ、N-bestを出力する
- (2) N-bestの中でrerankingした評価が最大になるように λ を最適化する
- (3) (1)~(2)を収束するまで繰り返す

(2)の最適化:

- Downhill simplex
- 各次元で順番に最適化(N-bestで離散化できる [Och 2003])
- Powell法

MERTの図解

development データ(原言語)

He ate a cake.
 She ate a soup.
 ...

decoder ← $\lambda_{LM} = 0.3, \lambda_{TM} = 0.3, \lambda_{RM} = 0.3$

各モデルの確率 (対数)付N-best出力

$f =$	LM	TM	RM	Total Cost
彼はケーキを食べ。	-3.8	-3.0	-8.1	-4.47
ヘリウムがケーキ。	-3.2	-4.3	-7.7	-4.56
彼がケーキを食べた。	-2.9	-3.5	-8.9	-4.59

Total Cost: -4.04, -4.11, -3.99

development データ(参照訳)

彼はケーキを食べた。
 彼女はスープを飲んだ。
 ...

参照と比較すると、N-bestの中でどの候補がよいのか分かる(BLEUやWER)

更新した $\lambda_{LM} = 0.4, \lambda_{TM} = 0.3, \lambda_{RM} = 0.2$

BLEUやWERの改善

λ を更新して繰り返す

まとめ: 午後の実習 (英日方向のシステム)

- 前半(2時間強)
 - 訓練データの準備
 - 英日データの分離と前処理
 - 言語モデルの構築
 - 50万文の日本語テキストと SRILMで5-gramモデルを作る
 - フレーズ翻訳モデルの構築と確認
 - Mosesのスク립トでフレーズ翻訳モデルを作成
 - 途中ファイルを確認
 - 方向のある単語アラインメント (IBMモデルの結果)
 - 対称化単語アラインメント
 - 抽出されたフレーズテーブル
 - Moses設定ファイルの作成
- 後半(1時間強)
 - チューニング
 - MERTの実行
 - 翻訳実験と評価
 - Mosesでテスト文(30文)の翻訳
 - BLEU値の計算
 - 最終課題
 - 指定された条件で全体を繰り返す

おまけ: SMTの研究テーマ

2nd Workshop on SMT 2007:
 Topics of interest include, but are not limited to:

- word-based, phrase-based, syntax-based SMT
- using comparable corpora for SMT
- using morphological and POS information for SMT
- integration of rule-based MT and statistical MT
- decoding
- error analysis
- evaluation techniques for MT

- 言語モデル
 - 音声認識と様子が違う&あまり研究されていないので穴場?
- 翻訳モデル
 - フレーズモデル推定の脱ヒューリスティックス
 - フレーズモデルの次: 構文情報をいかに取り込むか
 - Reordering Model (特に日英)
- デコーダ
 - Multistack & Beam-search がベスト?
 - 普通に考えると局所探索がよいに決まっているのだが...
- 自動評価
 - ルールベースとSMTが比較できる指標
- チューニング
- 対訳データ

しかし、いずれを研究するとしても大規模な日英対訳データがないことには面白くない?
 → NTCIR-7のデータを使いましょう!