

Khafra デコーダ 使用マニュアル

ver. 1.0

越川 満, 安田 隆浩, 乾 孝司, 山本 幹雄

自然言語処理 on the web 研究室 @ 筑波大学

1 Khafra デコーダとは

Khafra は筑波大学 自然言語処理 on the Web 研究室で開発したフレーズベース統計的機械翻訳デコーダです。[Zens and Ney 2008] にて提案された、「動的計画法とマルチスタックビームサーチ法を組み合わせた手法」を用いて翻訳文の探索を行います。これにより、Khafra では Moses[Koehn et al. 2007] より遥かに高速な翻訳を行うことができます。

2 インストール

2.1 各種ツールのインストール

Khafra を使用するために以下のものを用意します。各ツールのインストール方法はそれぞれのマニュアルに従ってください。

Moses 入手先 <http://www.statmt.org/moses/>

SRI Language Modeling Toolkit 入手先 <http://www.speech.sri.com/projects/srilm>

GIZA++,mkcls 入手先 <http://code.google.com/p/giza-pp/>

C++ コンパイラ

2.1.1 Khafra デコーダのインストール

Khafra デコーダをダウンロードします。

```
% wget http://www.nlp.mibel.cs.tsukuba.ac.jp/khafra/khafra_v*.**.tgz
```

ダウンロードしたファイルを展開します。

```
% tar -zxvf khafra_v*.**.tgz
```

展開されたディレクトリ khafra_v*.**に移動します。

```
% cd khafra_v*.**.*/
```

khafra_v*.**内にある Makefile を以下のように書き換えます。

```
% vi Makefile
```

- 1 行目 : SRILM=“SRI LM をインストールしたディレクトリへの絶対パス”
- 3 行目 : SRILM_LIB_DIR=\$SRILM/lib***への絶対パス (***)は環境により異なる)

khafra_v*.**ディレクトリ直下で、make コマンド (引数なし) を実行します。

```
% make
```

make により ./bin/khafra が作成されればインストール完了です。

```
% ls bin
```

khafra <—これが表示されればインストール完了

3 モデルの学習とチューニング

SRI Language Model Toolkit と Moses を使用して学習した言語モデル、フレーズ翻訳モデル、語順並べ替えモデル (msd-bidirectional-fe) を用意します。チューニングには Moses を使用して下さい。

4 サンプルデータ

4.1 テキストデータ

khafra/sample/text/以下にはサンプル用のテキストデータが入っています。

`train.{ab,12}` 翻訳モデル学習用の 1 万文の対訳文。

`train_lm.12` 言語モデル学習用の言語 12 にの文。 `train.12` と同じ。

`dev.{ab,12}` パラメータのチューニング用の対訳文。 `train` の最初の 1000 文。

`test.{ab,12}` テスト用の対訳 10 文。学習データとの重複はない。

サンプルテキストは、単語 “a” と単語 “b” からなる言語 `ab` と、単語 “1” と単語 “2” からなる言語 `12` との対訳データとなっています。なお、言語 `ab` と言語 `12` は、単語 “a” と単語 “1”、単語 “b” と単語 “2” がそれぞれ対応してます。また、言語 `12` は単語 “2” の後に単語 “1” が出現することはないという特徴を持ちます。 `train.{ab,12}` の内容は、以下のようにになっています。

`train.ab`

```
b a b a b b a b b a b b a b a a a a a
a b b b a b b b b a a b a
a b a b b a b a a b b a b a a a a b b
...
```

`train.12`

```
1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
1 1 1 1 1 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
...
```

4.2 モデルと設定ファイル

khafra/sample/model には、サンプルデータ train.{ab,12} および train_lm.12 から学習した、以下のファイルが含まれています。

lm.12 言語モデル

phrase-table フレーズ翻訳モデル

reordering-table 語順並べ替えモデル

moses.ini dev.{ab,12} に対してチューニング済みのパラメータを含む設定ファイル

5 Khafra の実行

5.1 実行に際しての注意事項

Khafra は、Moses で使用するものと同じフォーマットのモデル・設定ファイルを利用して翻訳を行います。翻訳に必要なファイルは、以下のとおりです。

- 言語モデル
- フレーズ翻訳モデル
- 語順並べ替えモデル
- 設定ファイル

なお、フレーズ翻訳モデルと語順並べ替えモデルのファイルが圧縮されている場合は、必ずファイルを展開してください。また、設定ファイルに記述されている各モデルへのパスは、「絶対パス」もしくは「khafra 実行時のカレントディレクトリからの相対パス」に変更してください。

5.2 サンプルデータを用いた実行例

サンプルモデルと設定ファイルを使用し、サンプルテキスト test.ab を翻訳する場合を説明します。まず、khafra のインストールディレクトリへ移動します。

```
% cd ***/khafra_v*.**
```

以下のコマンドを実行します。

```
$ bin/khakra -f sample/model/moses.ini < sample/text/test.ab > test.tlt
```

-f sample/model/moses.ini 設定ファイルにサンプル用の設定ファイルを指定
-i sample/text/test.ab サンプルテキストの test.ab を翻訳

翻訳結果が test.tlt に出力されます。なお、正解データは test.ab.ref です。

```
test.ab
a a b a b b b b
a b a b b b a b b
b b b a b a a b a a
...
```

```
test.ab.ref
1 1 1 2 2 2 2 2
1 1 1 2 2 2 2 2 2
1 1 1 1 1 1 2 2 2 2
...
```

6 オプション引数 一覧

オプション	引数	説明
-beam-threshold(-b)	閾値 (default: 10^{-5})	threshold pruning のための閾値を指定
-config(-f)	設定ファイルへのパス	設定ファイル (ini ファイル) を指定
-input(-i)	原言語ファイル	翻訳したいテキストファイルを指定
-stack(-s)	ビーム幅 (default: 200)	cardinality stack のビーム幅を指定
-ttable-limit(-ttl)	フレーズ数 (default: 20)	同じ原言語フレーズに対して考慮する 目的言語フレーズ候補の最大数を指定
-verbose(-v)	レベル (default: 0)	出力の詳細レベル (0~ 3) を指定
-use-future-dist-cost	-	future スコアに最小歪みコスト (見積り) を加算

参考文献

- [Zens and Ney 2008] R.Zens and H.Ney. 2008 “Improvements in dynamic programming beam search for phrase-based statistical machine translation.” In *Proceedings of IWSLT*. pp.198–205.
- [Koehn et al. 2007] P.Koehn and et al. 2007 “Moses: Open source toolkit for statistical machine translation.” In *Proceedings of the ACL Demo and Poster Session*, pp.177–180.
- [Stolcke 2002] A.Stolcke. 2002. “SRILM - an Extensible Language Modeling Toolkit.” In *Proceedings of International Conference on Spoken Language Processing*, pages 901–904.
- [安田ほか. 2010] 安田 隆浩, 越川 満, 乾 孝司, 山本 幹雄. 2010. 「Khafra:語順並べ替えモデルに対応した動的計画法に基づく SMT デコーダ」言語処理学会第 16 回年次大会発表予稿集, to appear.