＜コーパスの説明＞

日英翻訳エンジン学習・評価用対訳コーパス

本コーパスの内容は、翻訳機器学習用データ：20,000 文、評価用データ：1,500 文（日英対訳文）から構成されている。利用法は、各研究機関において、独自に開発した機械翻訳手法の改良を行う際､その精度や性能が以前に比べて性能が向上したかを確認できる標準的な学習用データと評価用データを提供するものである。

また、本コーパスは NICT 内にあるオンライン評価サーバにおいて使うことによりこのサーバにアクセスできれば翻訳精度の評価を遠隔地から行う事ができる。

本コーパスは International Workshop on Spoken Language Translation (略称 IWSLT) の 2005 年評価キャンペーンの日英翻訳で使用された基本旅行会話データセット (http://www.is.cs.cmu.edu/ iwslt2005)に基づいて作られたコーパスである。IWSLT とは、毎年開催される音声翻訳の公開評価キャンペーンで､研究ワークショップも同時に行われる｡その目的は、共同作業を促進し、学術交流を深めることにある。

現在リリースされているデータファイルの構成及びフォーマットに関する詳細は以下の通り。

文字コード

- 日本語: UTF-8
- 英語:　UTF-8

データセット

- IWSLT_CSTAR03 - IWSLT 2004 の開発セット (BTEC Task)
- IWSLT_IWSLT04 - IWSLT 2004 の評価セット (BTEC Task)
- IWSLT_IWSLT05 - IWSLT 2005 の評価セット (BTEC Task)

データファイル

IWSLT/Japanese-English

(トレーニングデータ)

+ train/TXT/IWSLT_2005.train.ja.txt	(日本語テキスト)
+ train/TXT/IWSLT_2005.train.en.txt	(英語テキスト)

(テストデータ)

+ test/TXT/IWSLT_CSTAR03.ja.txt	(日本語入力文)
+ test/TXT/IWSLT_CSTAR03.en.txt	(英語参照訳：一文)
+ test/TXT/IWSLT_CSTAR03.mref.en.txt	(英語参照訳：複数文)
+ test/TXT/IWSLT_IWSLT04.ja.txt	(日本語入力文)
+ test/TXT/IWSLT_IWSLT04.en.txt	(英語参照訳：一文)
+ test/TXT/IWSLT_IWSLT04.mref.en.txt	(英語参照訳：複数文)
+ test/TXT/IWSLT_IWSLT05.ja.txt	(日本語入力文)
+ test/TXT/IWSLT_IWSLT05.en.txt	(英語参照訳：一文)
+ test/TXT/IWSLT_IWSLT05.mref.en.txt	(英語参照訳：複数文)

(評価データ：ケース・センシティブ、句読点あり)

+ test/SGM/IWSLT_CSTAR03.ja.case+punc.src.sgm (日本語原文 SGML file)
+ test/SGM/IWSLT_CSTAR03.en.case+punc.mref.sgm (英語参照訳 SGML file)
+ test/SGM/IWSLT_IWSLT04.ja.case+punc.src.sgm (日本語原文 SGML file)
+ test/SGM/IWSLT_IWSLT04.en.case+punc.mref.sgm (英語参照訳 SGML file)
+ test/SGM/IWSLT_IWSLT05.ja.case+punc.src.sgm (日本語原文 SGML file)
+ test/SGM/IWSLT_IWSLT05.en.case+punc.mref.sgm (English reference SGML file)

(評価データ：インケース・インセンシティブ、句読点なし)

+ test/SGM/IWSLT_CSTAR03.ja.no_case+no_punc.src.sgm (日本語原文 SGML file)
+ test/SGM/IWSLT_CSTAR03.en.no_case+no_punc.mref.sgm (英語参照訳 SGML file)
+ test/SGM/IWSLT_IWSLT04.ja.no_case+no_punc.src.sgm (日本語原文 SGML file)
+ test/SGM/IWSLT_IWSLT04.en.no_case+no_punc.mref.sgm (英語参照訳 SGML file)
+ test/SGM/IWSLT_IWSLT05.ja.no_case+no_punc.src.sgm (日本語原文 SGML file)
+ test/SGM/IWSLT_IWSLT05.en.no_case+no_punc.mref.sgm (英語参照訳 SGML file)

データフォーマット

[TXT データファイル]

+ 文毎にユニークの ID が付与されており、対応する翻訳例は同じ ID を持つ。
+ 対応するファイルにおける文章の並び順は同じ。
+ 一文の参照訳を収めたファイル：

(a)ファイル名: IWSLT_<TASK>.<LANG>.txt

> train/TXT/IWSLT_2005.train.en.txt
> test/TXT/IWSLT_IWSLT04.en.txt

(b)ファイルフォーマット: <SENTENCE_ID>\01\<TEXT>

> TRAIN_00001\01\This is the first training sentence.
> TRAIN_00002\01\This is the second training sentence.
> IWSLT_IWSLT04_001\01\This is the first test sentence.
> IWSLT_IWSLT04_001\01\This is the second test sentence.

+ 複数の参照訳を収めたファイル:

(a)ファイル名: IWSLT_<TASK>.mref.<LANG>.txt

> test/TXT/IWSLT_IWSLT05.mref.en.txt

(b) ファイルフォーマット: <SENTENCE_ID>\<PARAPHRASE_ID>\<TEXT>

> IWSLT_IWSLT05_001\01\This is the first reference translation of the first sentence ID
> IWSLT_IWSLT05_001\02\This is the second reference translation of the first sentence ID
> IWSLT_IWSLT05_001\03\...
> ...
> IWSLT_IWSLT05_001\01\This is the first reference translation of the second sentence ID
> IWSLT_IWSLT05_001\02\This is the second reference translation of the second sentence ID
> IWSLT_IWSLT05_001\03\...
> ...

[SGM データファイル]

+各データセットには、二種類のＳＧＭファイルが準備されている。

(a) ケース・センシティブ(case-sensitive) 句読点あり、分かち書きあり

（公式な評価仕様）

- test/SGM/IWSLT_<TASK>.case+punc.src.<SRC_LANG>.sgm
- test/SGM/IWSLT_<TASK>.case+punc.mref.<TRG_LANG>.sgm

(b) ケース・インセンシティブ(case-insensitive）句読点なし

（本サーバーにて追加した評価仕様）

- test/SGM/IWSLT_<TASK>.no_case+no_punc.src.<SRC_LANG>.sgm
- test/SGM/IWSLT_<TASK>.no_case+no_punc.mref.<TRG_LANG>.sgm

+ フォーマット: BLEU/NIST 評価プログラムで必要とされる SGML フォーマット

(a) SRC ファイル:

```
> <srcset setid="set" srclang="source">
> <DOC docid="document">
> <seg> ...</seg>
> ...
> </DOC>
> </srcset>
```

(b) REF ファイル:

```
> <refset setid="set" srclang="source" trglang="target">
> <DOC docid="document" sysid="r01">
> <seg> ...</seg>
> ...
> </DOC>
> </refset>
> ...
> <refset setid="set" srclang="source" trglang="target">
> <DOC docid="document" sysid="r16">
> <seg> ...</seg>
> ...
> </DOC>
> </refset>
```

(c) TST ファイル:

```
> <tstset setid="set" srclang="source" trglang="target">
> <DOC docid="document" sysid="system">
> <seg> ...</seg>
> ...
> </DOC>
> </tstset>
```