Cross-Lingual IR task

Group's ID:

BKYTR

List of Run ID(s):

1. BKJEBDDS
2. BKJEBDFU
3. BKJEBKFU
4. BKJEECFU
5. BKJEMTFU

* NTCIR-1 = NACSIS Test Collection 1

1 Overall Approach

1) What basic approach do you take to Cross-Lingual Retrieval?:
- Query Translation: YES
- Document Translation:
- Other (Please specify):

2 Query construction

1) Automatically or manually?:

BKJEBDDS,BKJEBDFU,BKJEBKFU,BKJEMTFU: Automatic.

2) (If manually) query builder?:
- Domain expert:
- Computer system expert:
- Other (Please specify):

3) (If manually) To what degree is his(her) ability of understanding Japanese?:
- native speaker (Japanese):
- Using dictionaries, he/she can write an academic paper in Japanese language:
- Using dictionaries, he/she can read an academic paper in Japanese language:
- He/She had been learned Japanese language more than three months:
- He/She can't understand Japanese language:
- Other (Please specify):

4) (If manually) To what degree is his(her) ability of understanding English?:
- native speaker (English):
- Using dictionaries, he/she can write an academic paper in English:
- Using dictionaries, he/she can read an academic paper in English:
- He/She had been learned English more than three months:
- He/She can't understand English:
- Other (Please specify):

5) Average time to do complete query construction [in minutes]): 5/60

6) Method(s) used in constructing queries
- Tokenizing (uni-gram, bi-gram, n-gram, word, phrase, other): word
- Phrase identification from topics?: NO
- Syntactic parsing?: NO
- Word sense disambiguation?: NO
- Proper noun identification?: NO
- Automatic query expansion?: NO
  * Lexical resources such as thesaurus?:
  * Automatic relevance feedback?:
    + Local context analysis:
    + Other(s) (Please specify):
  * Other(s) (Please specify):
- Automatic addition of Boolean/proximity operators?: NO
- Other(s) (Please specify):

7) Spelling checking (including manual checking)?: NO

8) Correcting them?: NO

3 Methods used in query translation

1) Multilingual dictionary
- Externally-constructed one(s): NONE
  * Name:
  * Size [in entries] [in MB]:

- Internally-constructed one(s)
  * Source, material, construction method):

    We extracted the Japanese and English keyword fields from the documents in the ntc1-je0 collection. The Japanese keywords are paired with the English keywords from the same document in the order in which they appear in the keyword fields. When there are more than one English translations for the same Japanese keyword, the most frequent English translation found in the ntc1-je0 collection is selected as the translaton of a Japanese keyword.

  * Size [in entries] [in MB]: 373,447 entries in 23 MB

2) Corpus: NONE
- Parallel corpus
- Comparable corpus

3) Machine translation system (run BKJEMTFU)
- Externally-constructed system:
  * Name: GISELLE (a research system from University of Southern California Information Sciences Institute) uses phonetic translation, but does not have a technical term dictionary
  * Size [in entries] [in MB]: unknown

- Internally-constructed system
  * Features, etc.:
  * Size [in entries] [in MB]:

4) Other(s) (Please specify):

5) Manual effort involved in translation?: NO

6) Query expansion:
- Before query translation: NO
- After query translation: NO
- No query expansion: YES

7) Methods used in query expansion: NONE
- Automatic relevance feedback
- Automatic relevance feedback (local context analysis):
- Global relevance feedback:
- Thesaurus, lexicon, etc.:
- Other (Please specify):

8) Disambiguation when translating?: NO for dictionary based systems, unknown for machine translation

4 Searching

4.1 Search times

1) Run ID: BKJEBDDS

2) Computer time to search [average per query, in CPU seconds]: 2.0

1) Run ID: BKJEBDFU

2) Computer time to search [average per query, in CPU seconds]: 3.05

1) Run ID: BKJEBKFU
2) Computer time to search [average per query, in CPU seconds]: 3.05

1) Run ID: BKJEECFU
2) Computer time to search [average per query, in CPU seconds]: 1.9

1) Run ID: BKJEMTFU
2) Computer time to search [average per query, in CPU seconds]: 3.5

4.2 Searching methods

1) Vector space model?:

2) Probabilistic model?: YES

3) Other (Please specify):

4.3 Factors in ranking

1) TF (Term Frequency)?: YES

2) IDF (Inverse document frequency)?: YES

3) Other term weights? (Please specify):

4) Semantic closeness?: NO

5) Positional information in the document?: NO

6) Syntactic clues? NO

7) Proximity of terms?: NO

8) Document length?: YES

9) Other (Please specify):

Query length, collection length.

4.4 Machine information

1) Machine type for the experiment: Sun UltraSPARC 2

2) Was the machine dedicated or shared?: shared

3) Amount of hard disk storage [in MB]: 40 GB

4) Amount of RAM [in MB]: 512 MB

5) Clock rate of CPU [in MHz]: 148

4.5 Others

1) Brief description of features of your system not answered above:

2) Others (Please specify):

3)Your group has:
- Japanese native speaker(s): YES
- Member(s) who can understand Japanese language: YES
- No member who can understand Japanese language: NO

言語横断検索タスク（Cross-Lingual IR task）

チーム略称：CRL

実行ID（複数ある場合はすべて）：CRL1 CRL2 CRL11

（以下、差し支えのない範囲で具体的に書いて下さい）

※［　］内は単位　※ NTCIR-1 ＝ NACSISテストコレクション１

１　全体的なアプローチ

(1)言語横断検索に用いた基本的なアプローチは何か?:
・検索式の翻訳:

２　検索式の作成

(1)検索式の作成は自動的か手動か：完全な自動

(2)少しでも手動の操作が含まれる場合、誰が検索式を作成したか？：該当せず
・分野の専門家：
・計算機システムの専門家：
・その他（具体的に）：

(3)少しでも手動の操作が含まれる場合、検索者の日本語の能力はどの程度か？：該当せず
・native speaker:
・辞書を使用すれば、日本語の論文を書ける:
・辞書を使用すれば、日本語の論文が読める:
・日本語を3ヶ月以上学んだことがある:
・全くできない:
・その他（具体的に）:

(4)少しでも手動の操作が含まれる場合、検索者の英語の能力はどの程度か？：該当せず
・native speaker:
・辞書を使用すれば、英語の論文を書ける:
・辞書を使用すれば、英語の論文が読める:
・英語を3ヶ月以上学んだことがある:
・全くできない:
・その他（具体的に）:

(5) 検索式の作成を完了するまでの時間（１課題当たりの平均時間［分］）：0.5

(6)検索式作成に使用した方法
・索引単位への分割：uni-gram, bi-gram, その他のn-gram, 単語、フレーズ、その他）:単語
・フレーズの抽出：せず
・構文解析：せず
・語義の曖昧性解消：せず
・固有名詞の識別：せず
・検索式の自動拡張：せず
－シソーラスなど既存のツール：せず
－自動レレバンスフィードバック：せず
＊ローカルコンテクストアナリシス：せず
＊その他（具体的に）：せず
－その他（具体的に）：せず
・ブール演算子や近接演算子などの自動的付与：せず
・その他（具体的に）：せず

(7)誤字脱字やスペルのチェック（手動も含む）は行なったか？：せず

(8)誤字脱字やスペルの修正は行なったか？：せず

３　検索式の翻訳方式

(1)多言語辞書
・既存のもの
－名称：EDR(CRL2のみ)
－サイズ［語数］[MB]：?
・独自に構築
－情報源/材料/構築の手法：je0のキーワード部分から取得(CRL1のみ)
－サイズ［語数］[MB]：?

(2)コーパス
・パラレルコーパス
・コンパラブルコーパス

(3)機械翻訳システム

使用せず
・既存のもの
－名称：
－サイズ［語数］[MB]：
・独自に構築
－情報源/材料/構築の手法：
－サイズ［語数］[MB]：

(4)その他（具体的に）：せず

(5) 翻訳に人手が介在したか（具体的に）：せず

(6) 検索式の拡張を
・検索式を翻訳する前に行なった。
・検索式を翻訳した後に行なった。
・行なわなかった。

(7) 検索式拡張の方式
・自動レレバンスフィードバック
・自動レレバンスフィードバック（ローカル コンテクスト アナリシス）
・グローバル レレバンスフィードバック
・同義語辞書・シソーラスなど
・その他（具体的に）

(8) 翻訳語の選定にあたり、あいまい性を解消する工夫をした（具体的に）：せず

４　検索

４．１　検索時間

(1)実行ID：CRL1,2

(2)検索時間（１検索式に対する平均CPU時間［秒］）：500

４．２　検索モデル

(1)ベクトル空間型を用いたか？：せず

(2)確率型を用いたか？：?

(3)その他（具体的に）：ロバートソンの式の亜流

４．３　ランクづけの要素

(1)TF（語の出現頻度）：使用せず

(2)IDF：用いた

(3)その他の重みづけ（具体的に）：せず

(4)意味の近さ：せず

(5)文書中の位置：せず

(6)構文的な手がかり：せず

(7)語の近接（距離）：せず

(8)文書の長さ：せず

(9)その他（具体的に）：せず

４．４　計算機についての情報

(1)実験に使用した計算機：Sun Ultra 10

(2)その計算機は専用か共用か：専用

(3)ハードディスクの総容量 [GB] ：60

(4)RAMの総容量 [MB] ：1G

(5)CPUのクロック数 [MHz] ：?

４．５　その他

(1)上の質問で回答していないシステムの特色：なし

(2)その他（具体的に）：

今回の検索に，jeのデータから知識獲得してその知識を用いて
e データを検索するのは，若干クローズドの意味合いがでてくる
ように思います．

例えば，jeのデータをまるごと知識として扱ってよい場合，
je に対して検索を行ないその結果の中で eテキストに該当する
ものを取り出すという手法も考えられます．この場合
クロスリンガル検索といえなくなってきます．
(モノリンガル検索と同程度の精度が出てしまいます．)

je のデータを検索の知識ベース作成に用いることに関して
なんらかの対応が必要に思います．

(3)チームの構成員に：
・日本語のnative speakerがいる：いる
・日本語のわかる人がいる：いる
・日本語のわかる人はいない：該当せず

言語横断検索タスク（Cross-Lingual IR task）

チーム略称：FLAB1

実行ID（複数ある場合はすべて）：FLAB11, FLAB12 （手法は共通）

（以下、差し支えのない範囲で具体的に書いて下さい）

※［ ］内は単位　※ NTCIR-1 ＝ NACSISテストコレクション１

１　全体的なアプローチ

(1)言語横断検索に用いた基本的なアプローチは何か?:
・検索式の翻訳： ○
・文書の翻訳:
・その他（具体的に）:

２　検索式の作成

(1)検索式の作成は自動的か手動か：手動

(2)少しでも手動の操作が含まれる場合、誰が検索式を作成したか？：
・分野の専門家：
・計算機システムの専門家：
・その他（具体的に）：

(3)少しでも手動の操作が含まれる場合、検索者の日本語の能力はどの程度か？：
・native speaker: ○
・辞書を使用すれば、日本語の論文を書ける:
・辞書を使用すれば、日本語の論文が読める:
・日本語を3ヶ月以上学んだことがある:
・全くできない:
・その他（具体的に）:

(4)少しでも手動の操作が含まれる場合、検索者の英語の能力はどの程度か？：
・native speaker:
・辞書を使用すれば、英語の論文を書ける: ○
・辞書を使用すれば、英語の論文が読める:
・英語を3ヶ月以上学んだことがある:
・全くできない:
・その他（具体的に）:

(5) 検索式の作成を完了するまでの時間（１課題当たりの平均時間［分］）：

(6)検索式作成に使用した方法
・索引単位への分割：uni-gram, bi-gram, その他のn-gram, 単語、フレーズ、その他）:
・フレーズの抽出：
・構文解析：
・語義の曖昧性解消：
・固有名詞の識別：
・検索式の自動拡張：
－シソーラスなど既存のツール：
－自動レレバンスフィードバック：
＊ローカルコンテクストアナリシス：
＊その他（具体的に）：
－その他（具体的に）：
・ブール演算子や近接演算子などの自動的付与：
・その他（具体的に）：

(7)誤字脱字やスペルのチェック（手動も含む）は行なったか？：

(8)誤字脱字やスペルの修正は行なったか？：

３　検索式の翻訳方式

(1)多言語辞書
・既存のもの
－名称：
－サイズ［語数］［MB］：
・独自に構築　○
－情報源/材料/構築の手法：
－サイズ［語数］［MB］：

(2)コーパス
・パラレルコーパス
・コンパラブルコーパス

(3)機械翻訳システム
・既存のもの
－名称：
－サイズ［語数］［MB］：
・独自に構築
－情報源/材料/構築の手法：
－サイズ［語数］［MB］：

(4)その他（具体的に）：

(5) 翻訳に人手が介在したか（具体的に）：

(6) 検索式の拡張を
・検索式を翻訳する前に行なった。
・検索式を翻訳した後に行なった。
・行なわなかった。

(7) 検索式拡張の方式
・自動レレバンスフィードバック
・自動レレバンスフィードバック（ローカル コンテクスト アナリシス）
・グローバル レレバンスフィードバック
・同義語辞書・シソーラスなど
・その他（具体的に）

(8) 翻訳語の選定にあたり、あいまい性を解消する工夫をした（具体的に）：

４　検索

４．１　検索時間

(1)実行ID：

(2)検索時間（１検索式に対する平均CPU時間［秒］）：

４．２　検索モデル

(1)ベクトル空間型を用いたか？：

(2)確率型を用いたか？：

(3)その他（具体的に）：

４．３　ランクづけの要素

(1)TF（語の出現頻度）：

(2)IDF：

(3)その他の重みづけ（具体的に）：

(4)意味の近さ：

(5)文書中の位置：

(6)構文的な手がかり：

(7)語の近接（距離）：

(8)文書の長さ：

(9)その他（具体的に）：

４．４　計算機についての情報

(1)実験に使用した計算機：

(2)その計算機は専用か共用か：

(3)ハードディスクの総容量 [GB] ：

(4)RAMの総容量 [MB] ：

(5)CPUのクロック数 [MHz] ：

４．５　その他

(1)上の質問で回答していないシステムの特色：

(2)その他（具体的に）：

(3)チームの構成員に：
・日本語のnative speakerがいる：
・日本語のわかる人がいる：
・日本語のわかる人はいない：

言語横断検索タスク（Cross-Lingual IR task）

チーム略称：NTE15

実行ID（複数ある場合はすべて）：NTE153,NTE154

※［　］内は単位　※ NTCIR-1 ＝ NACSISテストコレクション１

１　全体的なアプローチ

(1)言語横断検索に用いた基本的なアプローチは何か?:
・検索式の翻訳:
・文書の翻訳:
・その他（具体的に）：対訳コーパスを利用し，類似度計算
(タームベクトルの内積)により日本語質問文から
英語検索式を自動生成

２　検索式の作成

(1)検索式の作成は自動的か手動か：自動

(2)少しでも手動の操作が含まれる場合、誰が検索式を作成したか？：

(3)少しでも手動の操作が含まれる場合、検索者の日本語の能力はどの程度か？：

(4)少しでも手動の操作が含まれる場合、検索者の英語の能力はどの程度か？：

(5) 検索式の作成を完了するまでの時間（１課題当たりの平均時間［分］）：0.3

(6)検索式作成に使用した方法
・索引単位への分割：uni-gram, bi-gram, その他のn-gram, 単語、フレーズ、その他）：
単語
・フレーズの抽出：NO
・構文解析：NO
・語義の曖昧性解消：NO
・固有名詞の識別：NO
・検索式の自動拡張：NO
・ブール演算子や近接演算子などの自動的付与：NO
・その他（具体的に）：

(7)誤字脱字やスペルのチェック（手動も含む）は行なったか？：NO

(8)誤字脱字やスペルの修正は行なったか？：NO

３　検索式の翻訳方式

(1)多言語辞書 … 利用しない

(2)コーパス
・パラレルコーパス … 利用(文書単位の対応づけのある対訳コーパス)
・コンパラブルコーパス

(3)機械翻訳システム … 利用しない

(4)その他（具体的に）：
日本語質問文から単語を切り出し，各日本語単語に対応する英単語を
取得する．英単語取得には，対訳コーパスから作成したタームベクトル
を利用し，ベクトルの内積値が高い単語を訳語として採用する．
得られた訳語をOR結合して検索式とする．

(5) 翻訳に人手が介在したか（具体的に）：NO

(6) 検索式の拡張を
・行なわなかった。

(7) 検索式拡張の方式

(8) 翻訳語の選定にあたり、あいまい性を解消する工夫をした（具体的に）：

４　検索

４．１　検索時間

(1)実行ID：NTE153,NTE154

(2)検索時間（１検索式に対する平均CPU時間［秒］）：0.98

４．２　検索モデル

(1)ベクトル空間型を用いたか？：YES

(2)確率型を用いたか？：NO

(3)その他（具体的に）：

４．３　ランクづけの要素

(1)TF（語の出現頻度）：YES

(2)IDF：YES

(3)その他の重みづけ（具体的に）：NO

(4)意味の近さ：NO

(5)文書中の位置：NO

(6)構文的な手がかり：NO

(7)語の近接（距離）：YES

(8)文書の長さ：YES

(9)その他（具体的に）：

４．４　計算機についての情報

(1)実験に使用した計算機：Sun SS-UA2

(2)その計算機は専用か共用か：共用

(3)ハードディスクの総容量 [GB] ：30

(4)RAMの総容量 [MB] ：1024

(5)CPUのクロック数 [MHz] ：296

４．５　その他

(1)上の質問で回答していないシステムの特色：

(2)その他（具体的に）：

(3)チームの構成員に：
・日本語のnative speakerがいる：YES
・日本語のわかる人がいる：YES
・日本語のわかる人はいない：NO

言語横断検索タスク（Cross-Lingual IR task）

チーム略称：SONIA

実行ID（複数ある場合はすべて）：SONIA1　SONIA2　SONIA3

（以下、差し支えのない範囲で具体的に書いて下さい）

※［　］内は単位　　※ NTCIR-1 ＝ NACSISテストコレクション１

１　全体的なアプローチ

(1) 言語横断検索に用いた基本的なアプローチは何か?:
・検索式の翻訳:

２　検索式の作成

(1) 検索式の作成は自動的か手動か：自動

(2) 少しでも手動の操作が含まれる場合、誰が検索式を作成したか？：
・分野の専門家：
・計算機システムの専門家：
・その他（具体的に）：

(3) 少しでも手動の操作が含まれる場合、検索者の日本語の能力はどの程度か？：
・native speaker:
・辞書を使用すれば、日本語の論文を書ける:
・辞書を使用すれば、日本語の論文が読める:
・日本語を3ヶ月以上学んだことがある:
・全くできない:
・その他（具体的に）:

(4) 少しでも手動の操作が含まれる場合、検索者の英語の能力はどの程度か？：
・native speaker:
・辞書を使用すれば、英語の論文を書ける:
・辞書を使用すれば、英語の論文が読める:
・英語を3ヶ月以上学んだことがある:
・全くできない:
・その他（具体的に）:

(5) 検索式の作成を完了するまでの時間（１課題当たりの平均時間［分］）：4分

(6) 検索式作成に使用した方法
・索引単位への分割：
<検索要求>の部分のみを辞書を用いて、日本語キーワードにまず分割
・検索式の自動拡張：
－その他（具体的に）：
NTC1-J0 から単語共起頻度をとって、相互情報的基準で閾値以上のものを、検索キーワードに追加
・その他（具体的に）：
最後に日本語キーワードを英語に翻訳

(7) 誤字脱字やスペルのチェック（手動も含む）は行なったか？：行なわない

(8) 誤字脱字やスペルの修正は行なったか？：行なわない

３　検索式の翻訳方式

(1) 多言語辞書
・独自に構築
－サイズ［語数］：46 万語

(2) コーパス
・パラレルコーパス

(3) 機械翻訳システム
・独自に構築
－情報源/材料/構築の手法：
文章を対象としたシステムではなく、辞書にある訳語候補の中から、入力された複数のキーワードの適した訳語を選択するもの

(4) その他（具体的に）：

(5) 翻訳に人手が介在したか（具体的に）：介在していない

(6) 検索式の拡張を
・検索式を翻訳する前に行なった。

(7) 検索式拡張の方式
・その他（具体的に）
予めコーパスから単語の共起頻度をとっておき、各検索キーワードに相互情報量的な基準で、閾値以上のものを、検索キーに追加。

(8) 翻訳語の選定にあたり、あいまい性を解消する工夫をした（具体的に）：
予め、ソース言語とターゲット言語でそれぞれ単語共起頻度データを計数する。可能な訳語候補の組合せのうち、それらの共起頻度を要素とするベクトルの方向が、日本語検索キーワード間の共起頻度を要素とするベクトルの方向に最も近付くような、訳語候補の組合せを選択する。

４　検索

４．１　検索時間

(1) 実行ID：SONIA1　SONIA2　SONIA3

(2) 検索時間（１検索式に対する平均CPU時間［秒］）：2.8 秒

４．２　検索モデル

(1) ベクトル空間型を用いたか？：はい

４．３　ランクづけの要素
tf-idf のみ

４．４　計算機についての情報

(1) 実験に使用した計算機：SUN UltraSPARC-II

(2) その計算機は専用か共用か：共用

(3) ハードディスクの総容量［GB］：50GB

(4) RAMの総容量［MB］：524MB

(5) CPUのクロック数［MHz］：296MHz

４．５　その他

(1) 上の質問で回答していないシステムの特色：該当データなし

(2) その他（具体的に）：該当データなし

(3) チームの構成員に：
・日本語のnative speakerがいる：

Cross-Lingual IR task

Group's ID: TSB

List of Run ID(s): TSB1, TSB2, TSB3, TSB4, TSB5, TSB6

1 Overall Approach

1) What basic approach do you take to Cross-Lingual Retrieval?:
- Query Translation:

2 Query construction

1) Automatically or manually?:
TSB1: automatic
TSB2: automatic+local feedback
TSB3: manual
TSB4: manual+local feedback
TSB5: automatic
TSB6: automatic, with syntactic analysis

2) (If manually) query builder?:
- Other (Please specify): a researcher in IR(TSB3&4)

3) (If manually) To what degree is his(her) ability of understanding Japanese?:
- native speaker (Japanese): TSB3&4

4) (If manually) To what degree is his(her) ability of understanding English?:
- Other (Please specify): Japanese/English bilingual(TSB3&4)

5) Average time to do complete query construction [in minutes]):
TSB1: a few seconds
TSB2: N.A.
TSB3: about one minute
TSB4: N.A.
TSB5&6: 0.02

6) Method(s) used in constructing queries
- Tokenizing (uni-gram, bi-gram, n-gram, word, phrase, other):
- Syntactic parsing?:
- Word sense disambiguation?:
- Proper noun identification?:
- Automatic query expansion?:
* Automatic relevance feedback?:
+ Other(s) (Please specify): local feedback(TSB2&4)

7) Spelling checking (including manual checking)?: TSB3&4

8) Correcting them?: TSB3&4

3 Methods used in query translation

3) Machine translation system
- Externally-constructed system: TSB1&2
* Name: Toshiba ASTRANSAC
* Size [in entries] [in MB]: 140M entries
- Externally-constructed system: TSB5&6
* Name: Toshiba The Honyaku Professional V4.0
* Size [in entries] [in MB]: 140M entries 70MB

5) Manual effort involved in translation?: TSB3&4

6) Query expansion:
- After query translation: TSB2&4
- No query expansion: TSB1,3,5&6

7) Methods used in query expansion:
- Automatic relevance feedback(local feedback) TSB2&4

8) Disambiguation when translating?:
MT disambiguation(TSB1&2)

4 Searching

4.1 Search times

1) Run ID: TSB1, TSB2, TSB3, TSB4, TSB5, TSB6

2) Computer time to search [average per query, in CPU seconds]:
0.1 (TSB5&6)

4.2 Searching methods

1) Vector Space model?: TSB5&6
2) Probabilistic model?: TSB1-4

4.3 Factors in ranking

1) TF (Term Frequency)?: TSB1-6
2) IDF (Inverse document frequency)?: TSB1-6
3) Other term weights? (Please specify): BM25(TSB1-4)
5) Positional information in the document?: TSB5&6
6) Syntactic clues? TSB6
7) Proximity of term? TSB5&6
8) Document length?: TSB1-4

4.4 Machine information

1) Machine type for the experiment: SUN U2/U1/SS20 etc(TSB1-4)
EQUIUM(TSB5&6)
2) Was the machine dedicated or shared?: shared(TSB1-4)
dedicated(TSB5&6)
3) Amount of hard disk storage [in MB]: 25000MB(TSB1-4)
2GB+2GB(TSB5&6)
4) Amount of RAM [in MB]: 512(TSB1-4) 320(TSB5&6)
5) Clock rate of CPU [in MHz]: 296(TSB1-4) Pentium2 45m(TSB5&6)

4.5 Others

1) Brief description of features of your system not answered above:
- combination of morpheme matching and string matching
- stemming
3)Your group has:
- Japanese native speaker(s):

言語横断検索タスク（Cross-Lingual IR task）

チーム略称： ULIS

実行ID（複数ある場合はすべて）：
ULIS1 ULIS2 ULIS3 ULIS4 ULIS5 ULIS6 ULIS7 ULIS8 ULIS9 ULIS10
ULIS11 ULIS12 ULIS13 ULIS14 ULIS15 ULIS16 ULIS17 ULIS18 ULIS19

「ULIS1～9とULIS11～ULIS19」はJ-E検索結果
「ULIS10」はJ-J検索結果

1　全体的なアプローチ

(1)言語横断検索に用いた基本的なアプローチは何か?:
・検索式の翻訳:

まず、形態素解析によって「検索要求」中の内容語を抽出し、検索タームとする
次に、以下の3つのステップで検索タームを翻訳する
ステップ1: 専門用語辞書を用いて辞書引きする
ステップ2: ステップ1で失敗した「カタカナ」文字列を「翻字」する
ステップ3: ステップ1,2で失敗した単語を一般辞書を用いて辞書引きする

検索タームが複合語の場合は、辞書を用いて語基に分割しながら上記の
3ステップを行う

実行IDと翻訳方式の対応を以下に示す

ULIS1, ULIS4, ULIS7, ULIS11, ULIS14, ULIS17: ステップ1のみ
ULIS2, ULIS5, ULIS8, ULIS12, ULIS15, ULIS18: ステップ1,2のみ
ULIS3, ULIS6, ULIS9, ULIS13, ULIS16, ULIS19: ステップ1～3全て適用

2　検索式の作成

(1)検索式の作成は自動的か手動か：自動

(5) 検索式の作成を完了するまでの時間（1課題当たりの平均時間［分］）：
ULIS1, ULIS4, ULIS7, ULIS11, ULIS14, ULIS17: 00.17分
ULIS2, ULIS5, ULIS8, ULIS12, ULIS15, ULIS18: 0.18分
ULIS3, ULIS6, ULIS9, ULIS13, ULIS16, ULIS19: 0.18分

(6)検索式作成に使用した方法
・索引単位への分割：uni-gram, bi-gram, その他のn-gram, 単語、フレーズ、その他）：
単語

(7)誤字脱字やスペルのチェック（手動も含む）は行なったか？： No

(8)誤字脱字やスペルの修正は行なったか？： No

3　検索式の翻訳方式

(1)多言語辞書
・独自に構築
－情報源/材料/構築の手法： EDR専門用語辞書(情報処理)
2語基からなる複合語対訳に対して、日本語エントリを
分割し、語基対訳辞書を作成した
－サイズ［語数］[MB]： 2.4万語(0.50MB)

－情報源/材料/構築の手法： EDR日英対訳辞書
1語基エントリ(単純語)のみを抽出した
－サイズ［語数］[MB]： 26万語(7.4MB)

(4)その他（具体的に）： NACSISコレクションの英語抄録から抽出した英単語bi-gram

(5) 翻訳に人手が介在したか（具体的に）： No

(6) 検索式の拡張を
・行なわなかった。

(8) 翻訳語の選定にあたり、あいまい性を解消する工夫をした（具体的に）：

EDR専門用語辞書から抽出した語基の対応頻度とNACSISコレクションから抽出した
英語bi-gramを用いて訳語候補ごとに「確率スコア」を計算し、スコアが上位の
候補を検索に用いた

実行IDと訳語候補数との対応を以下に示す

ULIS1, ULIS2, ULIS3, ULIS11, ULIS12, ULIS13: 上位1訳語
ULIS4, ULIS5, ULIS6, ULIS14, ULIS15, ULIS16: 上位3訳語
ULIS7, ULIS8, ULIS9, ULIS17, ULIS18, ULIS19: 上位10訳語

4　検索

4．1　検索時間

実行IDと検索時間（1検索式に対する平均CPU時間［秒］）の対応を以下に示す

ULIS1　0.22
ULIS2　0.24
ULIS3　0.27
ULIS4　0.31
ULIS5　0.32
ULIS6　0.37
ULIS7　0.40
ULIS8　0.42
ULIS9　0.48
ULIS10　0.24
ULIS11　0.22
ULIS12　0.23
ULIS13　0.27
ULIS14　0.31
ULIS15　0.32
ULIS16　0.37
ULIS17　0.45
ULIS18　0.47
ULIS19　0.55

4．2　検索モデル

(1)ベクトル空間型を用いたか？： Yes

(2)確率型を用いたか？： No

(3)その他（具体的に）：

4．3　ランクづけの要素

(1)TF（語の出現頻度）： Yes

(2)IDF： Yes

その他、使用せず

4．4　計算機についての情報

(1)実験に使用した計算機： Pro2333DL

(2)その計算機は専用か共用か： 共用

(3)ハードディスクの総容量［GB］： 6.4GB

(4)RAMの総容量 [MB] : 192MB

(5)CPUのクロック数 [MHz] : 333MHz

４．５　その他

(1)上の質問で回答していないシステムの特色：

(a) 「検索式翻訳」と「検索エンジン」が完全に独立しているので、
メンテナンスや使用モジュールの切替えが容易である

(b) 日英／英日双方向の言語横断検索が可能

(2)その他（具体的に）：

実行ID「ULIS n」と「ULIS n + 10」の違い:
専門用語辞書を作成する際の日本語エントリの分割法が異なる

(3)チームの構成員に：
・日本語のnative speakerがいる： Yes

Cross-Lingual IR task

Group's ID:

UMD

List of Run ID(s):

umd1

(Please answer questions below.)

* NTCIR-1 = NACSIS Test Collection 1

1 Overall Approach

1) What basic approach do you take to Cross-Lingual Retrieval?:
- Query Translation:
- Document Translation:
- Other (Please specify):

Query Translation

2 Query construction

1) Automatically or manually?:

Automatically

2) (If manually) query builder?:
- Domain expert:
- Computer system expert:
- Other (Please specify):

3) (If manually) To what degree is his(her) ability of understanding Japanese?:
- native speaker (Japanese):
- Using dictionaries, he/she can write an academic paper in Japanese language:
- Using dictionaries, he/she can read an academic paper in Japanese language:
- He/She had been learned Japanese language more than three months:
- He/She can't understand Japanese language:
- Other (Please specify):

4) (If manually) To what degree is his(her) ability of understanding English?:
- native speaker (English):
- Using dictionaries, he/she can write an academic paper in English:
- Using dictionaries, he/she can read an academic paper in English:
- He/She had been learned English more than three months:
- He/She can't understand English:
- Other (Please specify):

5) Average time to do complete query construction [in minutes]):

0.5 minute per query

6) Method(s) used in constructing queries
- Tokenizing (uni-gram, bi-gram, n-gram, word, phrase, other):
- Phrase identification from topics?:
- Syntactic parsing?:
- Word sense disambiguation?:
- Proper noun identification?:
- Automatic query expansion?:
  * Lexical resources such as thesaurus?:
  * Automatic relevance feedback?:
    + Local context analysis:
    + Other(s) (Please specify):
  * Other(s) (Please specify):
- Automatic addition of Boolean/proximity operators?:
- Other(s) (Please specify):

Our query construction consists of three step:

- Fields extraction:
In this first step fileds of <topic> (query number) and <description> are extracted from the topics file. The resulted file is of the query numbers together with their descriptions;

- Segmentation
The above file is passed to JUMAN2.2 for processing. The output file of JUMAN is processed again so that the result is the query number together with their segmented descriptions;

- Query translation
Query file created in step 2 is passed to our DQT system for translation. The result is queries together with their number in a format accepted by our search system INQUERY;

7) Spelling checking (including manual checking)?:

Our DQT automatically translates the queries, so no spelling checking is needed.

8) Correcting them?:

Not applicable

3 Methods used in query translation

1) Multilingual dictionary
- Externally-constructed one(s):
  * Name:
  * Size [in entries] [in MB]:
- Internally-constructed one(s)
  * Source, material, construction method):
  * Size [in entries] [in MB]:

We use a Japanese/English dictionary called "edict" freely available from Monash University.
(64433 entries, 2.6Mb)

2) Corpus
- Parallel corpus
- Comparable corpus

3) Machine translation system
- Externally-constructed system:
  * Name:
  * Size [in entries] [in MB]:
- Internally-constructed system
  * Features, etc.:
  * Size [in entries] [in MB]:

4) Other(s) (Please specify):

5) Manual effort involved in translation?:

No manual effort involved.

6) Query expansion:
- Before query translation:
- After query translation:
- No query expansion:

No query expansion used.

7) Methods used in query expansion:
- Automatic relevance feedback
- Automatic relevance feedback (local context analysis):
- Global relevance feedback:
- Thesaurus, lexicon, etc.:
- Other (Please specify):

8) Disambiguation when translating?:

No.

4 Searching

4.1 Search times

1) Run ID:

umd1

2) Computer time to search [average per query, in CPU seconds]:

As quickly as a 40 term query per second

4.2 Searching methods

1) Vector space model?:

2) Probabilistic model?:

INQUERY

3) Other (Please specify):

4.3 Factors in ranking

1) TF (Term Frequency)?:

Yes

2) IDF (Inverse document frequency)?:

Yes

3) Other term weights? (Please specify):

4) Semantic closeness?:

5) Positional information in the document?:

6) Syntactic clues?

7) Proximity of terms?:

8) Document length?:

Yes

9) Other (Please specify):

4.4 Machine information

1) Machine type for the experiment:

SPARC 20

2) Was the machine dedicated or shared?:

shared

3) Amount of hard disk storage [in MB]:

22GB

4) Amount of RAM [in MB]:

64MB

5) Clock rate of CPU [in MHz]:

33MHz

4.5 Others

1) Brief description of features of your system not answered above:

2) Others (Please specify):

3)Your group has:
- Japanese native speaker(s):
- Member(s) who can understand Japanese language:
- No member who can understand Japanese language:

None of the members in our group can understand Japanese. But we do have natice Japanese speakers available on an occassional basis.

言語横断検索タスク（Cross-Lingual IR task）

チーム略称：sstut

実行ID（複数ある場合はすべて）：sstut1,sstut2

（以下、差し支えのない範囲で具体的に書いて下さい）

※［ ］内は単位　※ NTCIR-1 ＝ NACSISテストコレクション１

１　全体的なアプローチ

(1)言語横断検索に用いた基本的なアプローチは何か?:
・検索式の翻訳:検索式を全部分文字列に分割し，多言語辞書(同義語を含む)を用いて翻訳した.
・文書の翻訳:していない.
・その他（具体的に）：

２　検索式の作成

(1)検索式の作成は自動的か手動か：自動.

(2)少しでも手動の操作が含まれる場合、誰が検索式を作成したか？：
・分野の専門家：
・計算機システムの専門家：
・その他（具体的に）：

(3)少しでも手動の操作が含まれる場合、検索者の日本語の能力はどの程度か？：
・native speaker:
・辞書を使用すれば、日本語の論文を書ける:
・辞書を使用すれば、日本語の論文が読める:
・日本語を3ヶ月以上学んだことがある:
・全くできない:
・その他（具体的に）：

(4)少しでも手動の操作が含まれる場合、検索者の英語の能力はどの程度か？：
・native speaker:
・辞書を使用すれば、英語の論文を書ける:
・辞書を使用すれば、英語の論文が読める:
・英語を3ヶ月以上学んだことがある:
・全くできない:
・その他（具体的に）：

(5) 検索式の作成を完了するまでの時間（１課題当たりの平均時間［分］）：0分.

(6)検索式作成に使用した方法
・索引単位への分割：uni-gram, bi-gram, その他のn-gram, 単語、フレーズ、その他）:n-gramフレーズ.
・フレーズの抽出：なし．ただし間接的に，抽出と同等の効果のある方法を使用した.
・構文解析：なし.
・語義の曖昧性解消：なし.
・固有名詞の識別：なし.
・検索式の自動拡張：同義語も含むように拡張した.
－シソーラスなど既存のツール：なし.
－自動レレバンスフィードバック：なし.
＊ローカルコンテクストアナリシス：なし.
＊その他（具体的に）：なし.
－その他（具体的に）：
・ブール演算子や近接演算子などの自動的付与：なし.
・その他（具体的に）：

(7)誤字脱字やスペルのチェック（手動も含む）は行なったか？：なし．ただし同等の効果がある方法を使用した.

(8)誤字脱字やスペルの修正は行なったか？：なし.

３　検索式の翻訳方式

(1)多言語辞書
・既存のもの
－名称：電気・電子・情報用語対訳辞典(日外アソシエーツ)
－サイズ 約79,000[語数] 2[MB]：
－名称：建築・土木用語対訳辞典(日外アソシエーツ)
－サイズ 約49,000[語数] 1[MB]：
－名称：コンピュータ用語辞典(日外アソシエーツ)
－サイズ 約27,000[語数] 0.7[MB]：
－名称：25万医学用語大辞典(日外アソシエーツ)
－サイズ 約480,000[語数] 6[MB]：
－名称：最新科学技術用語辞典(三修社)
－サイズ 約160,000[語数] 7[MB]：
・独自に構築
－情報源/材料/構築の手法：上の5つの辞書をつなげて一つの辞書として使用した.
－サイズ 579,116[語数] 17[MB]：

(2)コーパス
使用していない.

(3)機械翻訳システム
使用していない.

(4)その他（具体的に）：

(5) 翻訳に人手が介在したか（具体的に）：しない.

(6) 検索式の拡張を
・行なわなかった。ただし，翻訳の訳数を複数使用した.

(7) 検索式拡張の方式
・その他（具体的に）:同義語辞書と等価のものを使用した.

(8) 翻訳語の選定にあたり、あいまい性を解消する工夫をした（具体的に）：対象データが学会論文データベースであることを考慮し，技術用語の辞典を用いて翻訳した.

４　検索

４．１　検索時間

(1)実行ID：sstut1

(2)検索時間（１検索式に対する平均CPU時間［秒］）：5395.30[秒]

(1)実行ID：sstut2

(2)検索時間（１検索式に対する平均CPU時間［秒］）：13716.44[秒]

４．２　検索モデル

(1)ベクトル空間型を用いたか？：用いていない.

(2)確率型を用いたか？：用いていない.

(3)その他（具体的に）：dpマッチングアルゴリズムを拡張し，文字列単位で比較を行なう.

４．３　ランクづけの要素

(1)TF（語の出現頻度）：使用した.

(2)IDF：使用した.

(3)その他の重みづけ（具体的に）：なし.

(4)意味の近さ：使用せず.

(5)文書中の位置：使用せず.

(6)構文的な手がかり：使用せず.

(7)語の近接（距離）：使用せず.

(8)文書の長さ：使用せず.

(9)その他（具体的に）：それぞれの部分文字列についてtf,idfを計算し，tfが1の場合と，idfが0.05よりも大きい場合は有効な文字列ではないとしてスコアは0，それ以外は文字列の長さにidfを掛けたものをその部分文字列のスコアとする．各ドキュメントにおいてスコアの総和をとったものをそのドキュメントのスコアとし，それによってランク付けする．

４．４　計算機についての情報

(1)実験に使用した計算機：AT互換機

(2)その計算機は専用か共用か：共用.

(3)ハードディスクの総容量［GB］：8[GB]

(4)RAMの総容量［MB］：512[MB]

(5)CPUのクロック数［MHz］：400[MHz]

４．５　その他

(1)上の質問で回答していないシステムの特色：本システムの特色は検索式を単語に分割するということをせずに，全部分文字列に対して同義語も含めて英訳し検索するというものである．dpマッチングを使用した方法で，一つのベースラインを示す目的で参加した.

(2)その他（具体的に）：実行id=sstut1は<検索要求>のみを用いた検索結果.
実行id=sstut2は<検索要求>をのみ用いて得た検索結果の上位一万件のドキュメントをピックアップし，その後<検索要求>と<検索要求説明>を用いて検索したもの.

(3)チームの構成員に：
・日本語のnative speakerがいる：

```
Cross-Lingual IR task

Group's ID: TSTAR
List of Run ID(s): tstar1, tstar4, tstar7, tstar10, tstar21

(Please answer questions below.)
* NTCIR-1 = NACSIS Test Collection 1
1 Overall Approach
  1) What basic approach do you take to Cross-Lingual Retrieval?:   Query Translation
     - Query Translation:
     - Document Translation:
     - Other (Please specify):
2 Query construction
 1) Automatically or manually?:   Automatically
 2) (If manually) query builder?:
     - Domain expert:
     - Computer system expert:
     - Other (Please specify):
 3) (If manually) To what degree is his(her) ability of understanding Japanese?:
     - native speaker (Japanese):
     - Using dictionaries, he/she can write an academic paper in Japanese language:
     - Using dictionaries, he/she can read an academic paper in Japanese language:
     - He/She had been learned Japanese language more than three months:
     - He/She can't understand Japanese language:
     - Other (Please specify):
 4) (If manually) To what degree is his(her) ability of understanding English?:
     - native speaker (English):
     - Using dictionaries, he/she can write an academic paper in English:
     - Using dictionaries, he/she can read an academic paper in English:
     - He/She had been learned English more than three months:
     - He/She can't understand English:
     - Other (Please specify):
 5) Average time to do complete query construction [in minutes]):   0.01 minute
 6) Method(s) used in constructing queries
     - Tokenizing (uni-gram, bi-gram, n-gram, word, phrase, other): word
     - Phrase identification from topics?:  No.
     - Syntactic parsing?:  No.
     - Word sense disambiguation?:  No.
     - Proper noun identification?: No.
     - Automatic query expansion?:  No.
       * Lexical resources such as thesaurus?: No.
       * Automatic relevance feedback?: No.
          + Local context analysis:     No.
          + Other(s) (Please specify):  No.
       * Other(s) (Please specify):     No.
     - Automatic addition of Boolean/proximity operators?: No.
     - Other(s) (Please specify):
 7) Spelling checking (including manual checking)?: No.
 8) Correcting them?: No.
3 Methods used in query translation
 1) Multilingual dictionary
    - Externally-constructed one(s):
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed one(s)     Yes.
      * Source, material, construction method):
      * Size [in entries] [in MB]:      3 MB
 2) Corpus
    - Parallel corpus
    - Comparable corpus
 3) Machine translation system
    - Externally-constructed system:
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed system
      * Features, etc.:
      * Size [in entries] [in MB]:
 4) Other(s) (Please specify):
 5) Manual effort involved in translation?:   No.
```

```
 6) Query expansion:    No.
    - Before query translation:
    - After query translation:
    - No query expansion:
 7) Methods used in query expansion:    No.
    - Automatic relevance feedback
    - Automatic relevance feedback (local context analysis):
    - Global relevance feedback:
    - Thesaurus, lexicon, etc.:
    - Other (Please specify):
 8) Disambiguation when translating?:  No.
4 Searching
4.1 Search times
 1) Run ID:
 2) Computer time to search [average per query, in CPU seconds]:
        tstar1          4.71 seconds
        tstar4          6.22 seconds
        tstar7          5.66 seconds
        tstar10         6.22 seconds
        tstar21         8.86 seconds

4.2 Searching methods
 1) Vector space model?:  Yes
 2) Probabilistic model?:
 3) Other (Please specify):
4.3 Factors in ranking
  1) TF (Term Frequency)?:                      Yes
  2) IDF (Inverse document frequency)?:         Yes
  3) Other term weights? (Please specify):
  4) Semantic closeness?:
  5) Positional information in the document?:
  6) Syntactic clues?
  7) Proximity of terms?:
  8) Document length?:                          Yes
  9) Other (Please specify):
4.4 Machine information
  1) Machine type for the experiment:           Sun Spark Station 5
  2) Was the machine dedicated or shared?:      shared
  3) Amount of hard disk storage [in MB]:       8 GB
  4) Amount of RAM [in MB]:                     32 MB
  5) Clock rate of CPU [in MHz]:
4.5 Others
  1) Brief description of features of your system not answered above:
  2) Others (Please specify):
  3)Your group has:
     - Japanese native speaker(s):                              No.
     - Member(s) who can understand Japanese language:          One.
     - No member who can understand Japanese language:
```

```
Cross-Lingual IR task

Group's ID: TSTAR
List of Run ID(s): tstar2, tstar5, tstar8, tstar11, tstar22

(Please answer questions below.)
* NTCIR-1 = NACSIS Test Collection 1
1 Overall Approach
  1) What basic approach do you take to Cross-Lingual Retrieval?:   Query Translation
     - Query Translation:
     - Document Translation:
     - Other (Please specify):
2 Query construction
 1) Automatically or manually?:   Automatically
 2) (If manually) query builder?:
     - Domain expert:
     - Computer system expert:
     - Other (Please specify):
 3) (If manually) To what degree is his(her) ability of understanding Japanese?:
     - native speaker (Japanese):
     - Using dictionaries, he/she can write an academic paper in Japanese language:
     - Using dictionaries, he/she can read an academic paper in Japanese language:
     - He/She had been learned Japanese language more than three months:
     - He/She can't understand Japanese language:
     - Other (Please specify):
 4) (If manually) To what degree is his(her) ability of understanding English?:
     - native speaker (English):
     - Using dictionaries, he/she can write an academic paper in English:
     - Using dictionaries, he/she can read an academic paper in English:
     - He/She had been learned English more than three months:
     - He/She can't understand English:
     - Other (Please specify):
 5) Average time to do complete query construction [in minutes]):   0.01 minute
 6) Method(s) used in constructing queries
     - Tokenizing (uni-gram, bi-gram, n-gram, word, phrase, other): word
     - Phrase identification from topics?:  No.
     - Syntactic parsing?:  No.
     - Word sense disambiguation?:  No.
     - Proper noun identification?: No.
     - Automatic query expansion?:  No.
       * Lexical resources such as thesaurus?: No.
       * Automatic relevance feedback?: No.
          + Local context analysis:     No.
          + Other(s) (Please specify):  No.
       * Other(s) (Please specify):     No.
     - Automatic addition of Boolean/proximity operators?: No.
     - Other(s) (Please specify):
 7) Spelling checking (including manual checking)?: No.
 8) Correcting them?: No.
3 Methods used in query translation
 1) Multilingual dictionary
    - Externally-constructed one(s):
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed one(s)     Yes.
      * Source, material, construction method):
      * Size [in entries] [in MB]:      3 MB
 2) Corpus
    - Parallel corpus
    - Comparable corpus
    - Monolingual Corpus: LOB/NACSIS/TREC6 (Disc 4 and 5)
 3) Machine translation system
    - Externally-constructed system:
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed system
      * Features, etc.:
      * Size [in entries] [in MB]:
 4) Other(s) (Please specify):
 5) Manual effort involved in translation?:   No.
 6) Query expansion:    No.
    - Before query translation:
    - After query translation:
    - No query expansion:
 7) Methods used in query expansion:    No.
    - Automatic relevance feedback
    - Automatic relevance feedback (local context analysis):
    - Global relevance feedback:
    - Thesaurus, lexicon, etc.:
    - Other (Please specify):
 8) Disambiguation when translating?:  No.
4 Searching
4.1 Search times
 1) Run ID:
 2) Computer time to search [average per query, in CPU seconds]:
        tstar2          5.00 seconds
        tstar5          5.92 seconds
        tstar8          5.94 seconds
        tstar11         6.79 seconds
        tstar22         13.02 seconds

4.2 Searching methods
 1) Vector space model?:  Yes
 2) Probabilistic model?:
 3) Other (Please specify):
4.3 Factors in ranking
  1) TF (Term Frequency)?:                      Yes
  2) IDF (Inverse document frequency)?:         Yes
  3) Other term weights? (Please specify):
  4) Semantic closeness?:
  5) Positional information in the document?:
  6) Syntactic clues?
  7) Proximity of terms?:
  8) Document length?:                          Yes
  9) Other (Please specify):
4.4 Machine information
  1) Machine type for the experiment:           Sun Spark Station 5
  2) Was the machine dedicated or shared?:      shared
  3) Amount of hard disk storage [in MB]:       8 GB
  4) Amount of RAM [in MB]:                     32 MB
  5) Clock rate of CPU [in MHz]:
4.5 Others
  1) Brief description of features of your system not answered above:
  2) Others (Please specify):
  3)Your group has:
     - Japanese native speaker(s):                                  No.
     - Member(s) who can understand Japanese language:              One.
     - No member who can understand Japanese language:
```

```
Cross-Lingual IR task

Group's ID: TSTAR
List of Run ID(s): tstar3, tstar6, tstar9, tstar12

(Please answer questions below.)
* NTCIR-1 = NACSIS Test Collection 1
1 Overall Approach
  1) What basic approach do you take to Cross-Lingual Retrieval?:   Query Translation
     - Query Translation:
     - Document Translation:
     - Other (Please specify):
2 Query construction
 1) Automatically or manually?:   Automatically
 2) (If manually) query builder?:
     - Domain expert:
     - Computer system expert:
     - Other (Please specify):
 3) (If manually) To what degree is his(her) ability of understanding Japanese?:
     - native speaker (Japanese):
     - Using dictionaries, he/she can write an academic paper in Japanese language:
     - Using dictionaries, he/she can read an academic paper in Japanese language:
     - He/She had been learned Japanese language more than three months:
     - He/She can't understand Japanese language:
     - Other (Please specify):
 4) (If manually) To what degree is his(her) ability of understanding English?:
     - native speaker (English):
     - Using dictionaries, he/she can write an academic paper in English:
     - Using dictionaries, he/she can read an academic paper in English:
     - He/She had been learned English more than three months:
     - He/She can't understand English:
     - Other (Please specify):
 5) Average time to do complete query construction [in minutes]):   0.75 minute
 6) Method(s) used in constructing queries
     - Tokenizing (uni-gram, bi-gram, n-gram, word, phrase, other): word
     - Phrase identification from topics?:  No.
     - Syntactic parsing?:  No.
     - Word sense disambiguation?:  Yes. (word co-occurrence)
     - Proper noun identification?: No.
     - Automatic query expansion?:  Yes.
       * Lexical resources such as thesaurus?: No.
       * Automatic relevance feedback?: No.
          + Local context analysis:     No.
          + Other(s) (Please specify):  No.
       * Other(s) (Please specify):     word co-occurrence
     - Automatic addition of Boolean/proximity operators?: No.
     - Other(s) (Please specify):
 7) Spelling checking (including manual checking)?: No.
 8) Correcting them?: No.
3 Methods used in query translation
 1) Multilingual dictionary
    - Externally-constructed one(s):
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed one(s)     Yes.
      * Source, material, construction method):
      * Size [in entries] [in MB]:      3 MB
 2) Corpus
    - Parallel corpus
    - Comparable corpus
    - Monolingual Corpus:       LOB
 3) Machine translation system
    - Externally-constructed system:
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed system
      * Features, etc.:
      * Size [in entries] [in MB]:
 4) Other(s) (Please specify):
```



```
 5) Manual effort involved in translation?:   No.
 6) Query expansion:    Yes!
    - Before query translation:
    - After query translation:  Yes!
    - No query expansion:
 7) Methods used in query expansion:
    - Automatic relevance feedback
    - Automatic relevance feedback (local context analysis):
    - Global relevance feedback:
    - Thesaurus, lexicon, etc.:
    - Other (Please specify):   word co-occurrence
 8) Disambiguation when translating?:  Yes!
4 Searching
4.1 Search times
 1) Run ID:
 2) Computer time to search [average per query, in CPU seconds]:
        tstar3          3.00 seconds
        tstar6          5.92 seconds
        tstar9          5.94 seconds
        tstar12         6.79 seconds


4.2 Searching methods
 1) Vector space model?:  Yes
 2) Probabilistic model?:
 3) Other (Please specify):
4.3 Factors in ranking
  1) TF (Term Frequency)?:                      Yes
  2) IDF (Inverse document frequency)?:         Yes
  3) Other term weights? (Please specify):
  4) Semantic closeness?:
  5) Positional information in the document?:
  6) Syntactic clues?
  7) Proximity of terms?:
  8) Document length?:                          Yes
  9) Other (Please specify):
4.4 Machine information
  1) Machine type for the experiment:           Sun Spark Station 5
  2) Was the machine dedicated or shared?:      shared
  3) Amount of hard disk storage [in MB]:       8 GB
  4) Amount of RAM [in MB]:                     32 MB
  5) Clock rate of CPU [in MHz]:
4.5 Others
  1) Brief description of features of your system not answered above:
  2) Others (Please specify):
  3)Your group has:
     - Japanese native speaker(s):                                  No.
     - Member(s) who can understand Japanese language:              One.
     - No member who can understand Japanese language:
```

```
Cross-Lingual IR task

Group's ID: TSTAR
List of Run ID(s): tstar17, tstar18, tstar19, tstar20

(Please answer questions below.)
* NTCIR-1 = NACSIS Test Collection 1
1 Overall Approach
  1) What basic approach do you take to Cross-Lingual Retrieval?:   Query Translation
     - Query Translation:
     - Document Translation:
     - Other (Please specify):
2 Query construction
 1) Automatically or manually?:   Automatically
 2) (If manually) query builder?:
     - Domain expert:
     - Computer system expert:
     - Other (Please specify):
 3) (If manually) To what degree is his(her) ability of understanding Japanese?:
     - native speaker (Japanese):
     - Using dictionaries, he/she can write an academic paper in Japanese language:
     - Using dictionaries, he/she can read an academic paper in Japanese language:
     - He/She had been learned Japanese language more than three months:
     - He/She can't understand Japanese language:
     - Other (Please specify):
 4) (If manually) To what degree is his(her) ability of understanding English?:
     - native speaker (English):
     - Using dictionaries, he/she can write an academic paper in English:
     - Using dictionaries, he/she can read an academic paper in English:
     - He/She had been learned English more than three months:
     - He/She can't understand English:
     - Other (Please specify):
 5) Average time to do complete query construction [in minutes]):   0.25 minute
 6) Method(s) used in constructing queries
     - Tokenizing (uni-gram, bi-gram, n-gram, word, phrase, other): word
     - Phrase identification from topics?:  No.
     - Syntactic parsing?:  No.
     - Word sense disambiguation?:  Yes. (word co-occurrence)
     - Proper noun identification?: No.
     - Automatic query expansion?:  Yes.
       * Lexical resources such as thesaurus?: No.
       * Automatic relevance feedback?: No.
          + Local context analysis:     No.
          + Other(s) (Please specify):  No.
       * Other(s) (Please specify):     word co-occurrence
     - Automatic addition of Boolean/proximity operators?: No.
     - Other(s) (Please specify):
 7) Spelling checking (including manual checking)?: No.
 8) Correcting them?: No.
3 Methods used in query translation
 1) Multilingual dictionary
    - Externally-constructed one(s):
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed one(s)     Yes.
      * Source, material, construction method):
      * Size [in entries] [in MB]:      3 MB
 2) Corpus
    - Parallel corpus
    - Comparable corpus
    - Monolingual Corpus:       NACSIS E-Collection
 3) Machine translation system
    - Externally-constructed system:
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed system
      * Features, etc.:
      * Size [in entries] [in MB]:
 4) Other(s) (Please specify):
```

```
 5) Manual effort involved in translation?:   No.
 6) Query expansion:    Yes!
    - Before query translation:
    - After query translation:  Yes!
    - No query expansion:
 7) Methods used in query expansion:
    - Automatic relevance feedback
    - Automatic relevance feedback (local context analysis):
    - Global relevance feedback:
    - Thesaurus, lexicon, etc.:
    - Other (Please specify):   word co-occurrence
 8) Disambiguation when translating?:  Yes!
4 Searching
4.1 Search times
 1) Run ID:
 2) Computer time to search [average per query, in CPU seconds]:
        tstar17         4.96 seconds
        tstar18         6.28 seconds
        tstar19         5.66 seconds
        tstar20         6.28 seconds

4.2 Searching methods
 1) Vector space model?:  Yes
 2) Probabilistic model?:
 3) Other (Please specify):
4.3 Factors in ranking
  1) TF (Term Frequency)?:                      Yes
  2) IDF (Inverse document frequency)?:         Yes
  3) Other term weights? (Please specify):
  4) Semantic closeness?:
  5) Positional information in the document?:
  6) Syntactic clues?
  7) Proximity of terms?:
  8) Document length?:                          Yes
  9) Other (Please specify):
4.4 Machine information
  1) Machine type for the experiment:           Sun Spark Station 5
  2) Was the machine dedicated or shared?:      shared
  3) Amount of hard disk storage [in MB]:       8 GB
  4) Amount of RAM [in MB]:                     32 MB
  5) Clock rate of CPU [in MHz]:
4.5 Others
  1) Brief description of features of your system not answered above:
  2) Others (Please specify):
  3)Your group has:
     - Japanese native speaker(s):                              No.
     - Member(s) who can understand Japanese language:          One.
     - No member who can understand Japanese language:
```

```
Cross-Lingual IR task

Group's ID: TSTAR
List of Run ID(s): tstar23

(Please answer questions below.)
* NTCIR-1 = NACSIS Test Collection 1
1 Overall Approach
  1) What basic approach do you take to Cross-Lingual Retrieval?:   Query Translation
     - Query Translation:
     - Document Translation:
     - Other (Please specify):
2 Query construction
 1) Automatically or manually?:   Automatically
 2) (If manually) query builder?:
     - Domain expert:
     - Computer system expert:
     - Other (Please specify):
 3) (If manually) To what degree is his(her) ability of understanding Japanese?:
     - native speaker (Japanese):
     - Using dictionaries, he/she can write an academic paper in Japanese language:
     - Using dictionaries, he/she can read an academic paper in Japanese language:
     - He/She had been learned Japanese language more than three months:
     - He/She can't understand Japanese language:
     - Other (Please specify):
 4) (If manually) To what degree is his(her) ability of understanding English?:
     - native speaker (English):
     - Using dictionaries, he/she can write an academic paper in English:
     - Using dictionaries, he/she can read an academic paper in English:
     - He/She had been learned English more than three months:
     - He/She can't understand English:
     - Other (Please specify):
 5) Average time to do complete query construction [in minutes]):
               50 minutes  (total time for 53 queries)
               0.94 minute (per query)

 6) Method(s) used in constructing queries
     - Tokenizing (uni-gram, bi-gram, n-gram, word, phrase, other): word
     - Phrase identification from topics?:  No.
     - Syntactic parsing?:  No.
     - Word sense disambiguation?:  Yes. (word co-occurrence)
     - Proper noun identification?: No.
     - Automatic query expansion?:  Yes.
       * Lexical resources such as thesaurus?: No.
       * Automatic relevance feedback?: No.
          + Local context analysis:     No.
          + Other(s) (Please specify):  No.
       * Other(s) (Please specify):     word co-occurrence
     - Automatic addition of Boolean/proximity operators?: No.
     - Other(s) (Please specify):
 7) Spelling checking (including manual checking)?: No.
 8) Correcting them?: No.
3 Methods used in query translation
 1) Multilingual dictionary
    - Externally-constructed one(s):
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed one(s)     Yes.
      * Source, material, construction method):
      * Size [in entries] [in MB]:      3 MB
 2) Corpus
    - Parallel corpus
    - Comparable corpus
    - Monolingual Corpus:               TREC6 (Disc 4 and 5)
 3) Machine translation system
    - Externally-constructed system:
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed system
      * Features, etc.:
      * Size [in entries] [in MB]:
 4) Other(s) (Please specify):
 5) Manual effort involved in translation?:   No.
 6) Query expansion:    Yes!
    - Before query translation:
    - After query translation:  Yes!
    - No query expansion:
 7) Methods used in query expansion:
    - Automatic relevance feedback
    - Automatic relevance feedback (local context analysis):
    - Global relevance feedback:
    - Thesaurus, lexicon, etc.:
    - Other (Please specify):   word co-occurrence
 8) Disambiguation when translating?:  Yes!
4 Searching
4.1 Search times
 1) Run ID:
 2) Computer time to search [average per query, in CPU seconds]:
        tstar23          10.18 seconds

4.2 Searching methods
 1) Vector space model?:  Yes
 2) Probabilistic model?:
 3) Other (Please specify):
4.3 Factors in ranking
  1) TF (Term Frequency)?:                      Yes
  2) IDF (Inverse document frequency)?:         Yes
  3) Other term weights? (Please specify):
  4) Semantic closeness?:
  5) Positional information in the document?:
  6) Syntactic clues?
  7) Proximity of terms?:
  8) Document length?:                          Yes
  9) Other (Please specify):
4.4 Machine information
  1) Machine type for the experiment:           Sun Spark Station 5
  2) Was the machine dedicated or shared?:      shared
  3) Amount of hard disk storage [in MB]:       8 GB
  4) Amount of RAM [in MB]:                     32 MB
  5) Clock rate of CPU [in MHz]:
4.5 Others
  1) Brief description of features of your system not answered above:
  2) Others (Please specify):
  3)Your group has:
     - Japanese native speaker(s):                                  No.
     - Member(s) who can understand Japanese language:              One.
     - No member who can understand Japanese language:
```

```
Cross-Lingual IR task

Group's ID: TSTAR
List of Run ID(s): tstar13, tstar14, tstar15, tstar16

(Please answer questions below.)
* NTCIR-1 = NACSIS Test Collection 1
1 Overall Approach
  1) What basic approach do you take to Cross-Lingual Retrieval?:   Query Translation
     - Query Translation:
     - Document Translation:
     - Other (Please specify):
2 Query construction
 1) Automatically or manually?:   Automatically
 2) (If manually) query builder?:
     - Domain expert:
     - Computer system expert:
     - Other (Please specify):
 3) (If manually) To what degree is his(her) ability of understanding Japanese?:
     - native speaker (Japanese):
     - Using dictionaries, he/she can write an academic paper in Japanese language:
     - Using dictionaries, he/she can read an academic paper in Japanese language:
     - He/She had been learned Japanese language more than three months:
     - He/She can't understand Japanese language:
     - Other (Please specify):
 4) (If manually) To what degree is his(her) ability of understanding English?:
     - native speaker (English):
     - Using dictionaries, he/she can write an academic paper in English:
     - Using dictionaries, he/she can read an academic paper in English:
     - He/She had been learned English more than three months:
     - He/She can't understand English:
     - Other (Please specify):
 5) Average time to do complete query construction [in minutes]):
                  2 minutes  (total time for 53 queries)

 6) Method(s) used in constructing queries
     - Tokenizing (uni-gram, bi-gram, n-gram, word, phrase, other): word
     - Phrase identification from topics?:  No.
     - Syntactic parsing?:  No.
     - Word sense disambiguation?:  Yes. (word co-occurrence)
     - Proper noun identification?: No.
     - Automatic query expansion?:  Yes.
       * Lexical resources such as thesaurus?: No.
       * Automatic relevance feedback?: No.
          + Local context analysis:     No.
          + Other(s) (Please specify):  No.
       * Other(s) (Please specify):     word co-occurrence
     - Automatic addition of Boolean/proximity operators?: No.
     - Other(s) (Please specify):
 7) Spelling checking (including manual checking)?: No.
 8) Correcting them?: No.
3 Methods used in query translation
 1) Multilingual dictionary
    - Externally-constructed one(s):
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed one(s)     Yes.
      * Source, material, construction method):
      * Size [in entries] [in MB]:      3 MB
 2) Corpus
    - Parallel corpus
    - Comparable corpus
    - Monolingual Corpus:               TREC6 (Disc 4 and 5)
 3) Machine translation system
    - Externally-constructed system:
      * Name:
      * Size [in entries] [in MB]:
    - Internally-constructed system
      * Features, etc.:
      * Size [in entries] [in MB]:
 4) Other(s) (Please specify):
 5) Manual effort involved in translation?:   No.
 6) Query expansion:    Yes!
    - Before query translation:
    - After query translation:  Yes!
    - No query expansion:
 7) Methods used in query expansion:
    - Automatic relevance feedback
    - Automatic relevance feedback (local context analysis):
    - Global relevance feedback:
    - Thesaurus, lexicon, etc.:
    - Other (Please specify):   word co-occurrence
 8) Disambiguation when translating?:  Yes!
4 Searching
4.1 Search times
 1) Run ID:
 2) Computer time to search [average per query, in CPU seconds]:
        tstar13         5.09 seconds
        tstar14         6.60 seconds
        tstar15         5.86 seconds
        tstar16         6.45 seconds

4.2 Searching methods
 1) Vector space model?:  Yes
 2) Probabilistic model?:
 3) Other (Please specify):
4.3 Factors in ranking
  1) TF (Term Frequency)?:                      Yes
  2) IDF (Inverse document frequency)?:         Yes
  3) Other term weights? (Please specify):
  4) Semantic closeness?:
  5) Positional information in the document?:
  6) Syntactic clues?
  7) Proximity of terms?:
  8) Document length?:                          Yes
  9) Other (Please specify):
4.4 Machine information
  1) Machine type for the experiment:           Sun Spark Station 5
  2) Was the machine dedicated or shared?:      shared
  3) Amount of hard disk storage [in MB]:       8 GB
  4) Amount of RAM [in MB]:                     32 MB
  5) Clock rate of CPU [in MHz]:
4.5 Others
  1) Brief description of features of your system not answered above:
  2) Others (Please specify):
  3)Your group has:
     - Japanese native speaker(s):                                  No.
     - Member(s) who can understand Japanese language:              One.
     - No member who can understand Japanese language:
```

This directory contains the file lists of the results from the group TSTAR.

Because of the large size of the files (totally 46MB), we compress it
and upload to the ftp site. Excuse us that you have to uncompress it
using WinZIP or "pkunzip" on PC, or "unzip" on WorkStation.

1. tstar.zip
   The results of 23 runs (tstar1, tstar2, ..., tstar23).
   They are compressed by WinZip on PC.
   If you have problems to unzip the files, please let us know.

2. six system description files (tstar-*.txt)
   There are six system description files. Their cooresponding runs
   are listed as below:

tstar-sys-1st.txt	the 1st,4th,7th,10th,21st runs
tstar-sys-all.txt	the 2nd,5th,8th,11th,22nd runs
tstar-sys-lob.txt	the 3rd,6th,9th,12th runs
tstar-sys-trec6,txt	the 13th,14th,15th,16th runs
tstar-sys-nacsis.txt	the 17th,18th,19th,20th runs
tstar-sys-trec6-tdnc.txt	the 23rd run

Group TSTAR