NTCIR-9，NTCIR-10特許機械翻訳タスクでの人手評価

後藤功雄

情報通信研究機構

AAMT/Japio 特許文書の機械翻訳結果評価方法検討会 2012/09/07

発表の内容

- 特許機械翻訳タスクの概要
- 人手評価の必要性
- NTCIR-9で実施した人手評価
  - 文の訳質の評価（Adequacy/Acceptability）
  - 評価の検証と必要な文数の検討
- NTCIR-10での新しい人手評価
  - 特許審査評価
- まとめ

特許機械翻訳タスクの概要

- 主催者が特許翻訳用の訓練データとテストデータを用意
- 参加者が各自のシステムでテストデータを機械翻訳
- 主催者が翻訳結果を評価
- 参加者が研究成果をワークショップで発表

- 構築したデータ（訓練，テスト，評価結果，翻訳結果）は研究利用できるように管理

NTCIR-9:　実施済み（2010/10～2011/12）
NTCIR-10: 実施中　（2012/03～2013/06）

人手評価の必要性

- 自動評価
  - システム間の比較に有用
  - 課題
    - 自動評価は完璧ではない
      - RBMTとSMTなど手法が大きく異なるシステム間の比較は信頼性が低い
    - 自動評価値からはどれだけの訳質が達成できたかが明確ではない.
      - BLEU 20, 30, 40, 50
      - 翻訳元文の意味が分かる文の割合は？・・・不明

人手評価は最も信頼性が高い評価方法
特許機械翻訳タスクでは，人手評価をメインの評価として実施

NTCIR-9での評価の設計

- MTの用途の設定
  - 情報収集のための利用
    - 後編集を前提としない
    - 翻訳結果から原文の内容を理解したい
- 評価基準
  - Adequacy
    - 目的:システム間の比較
  - Acceptability
    - 目的:原文の意味が理解できる文数の割合を明らかにする

評価の実施方法

- 本評価
  - 各システムあたりランダムに選択した300文を評価
  - 評価者数：3人（有償）
  - 各評価者は各システムあたり100文を評価
  - 原文が同じである翻訳結果は同時に評価

- 事前のトレーニング
  - 翻訳結果100文を3人が評価
  - 3人で相談して統一の評価値を決定
  - この統一評価値を参照して本評価を実施

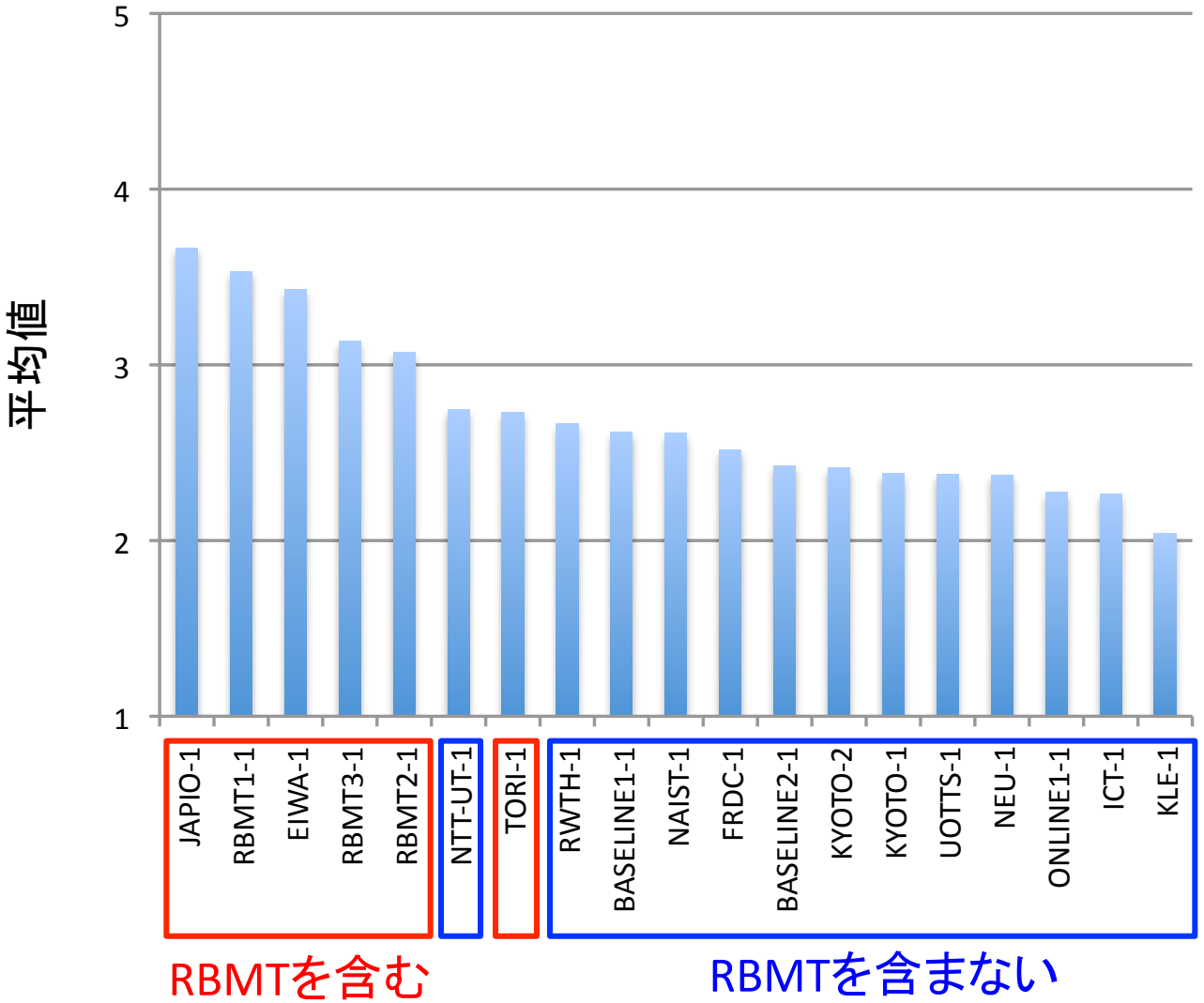
Adequacy

- 文単位の評価値
- 翻訳結果における翻訳元文中の意味の保持度合を5段階で評価

5	All
4	Most
3	Much
2	Little
1	None

- 評価レンジが広く，低い訳質から高い訳質まで分類できる.

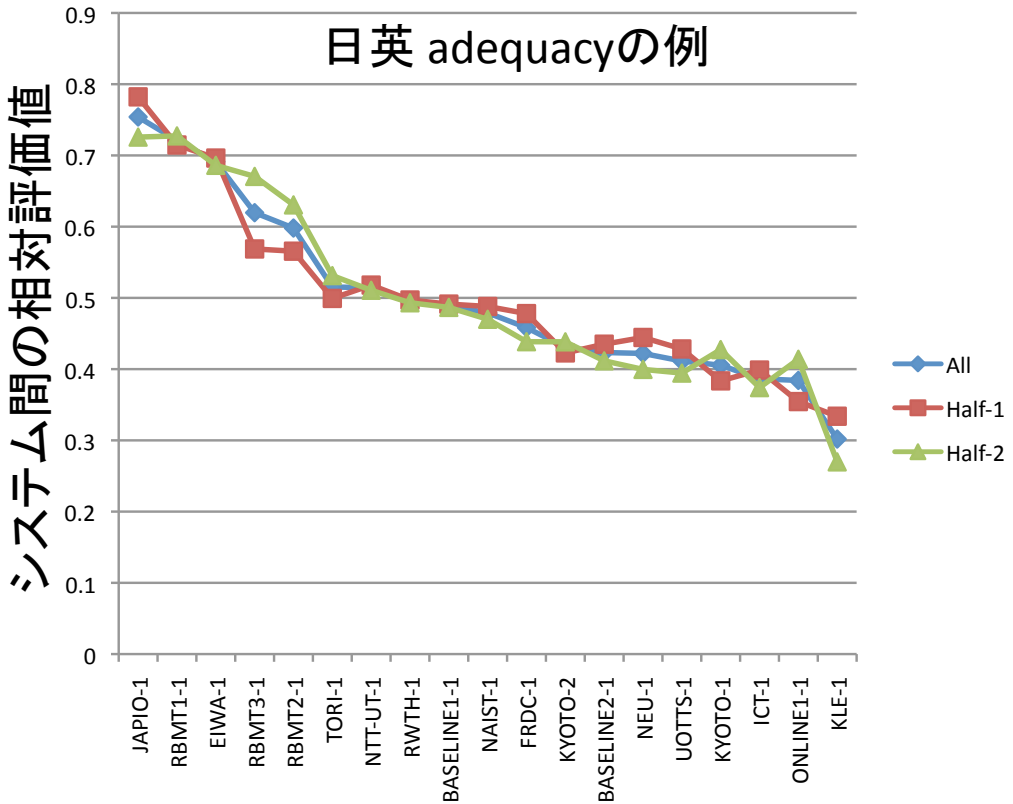
Adequacy評価の例（NTCIR-9 日英）



評価の検証と必要な文数の検討

- 評価文数（300文）はシステムの比較評価に十分か？
  - 前半150文と後半150文に分けて比較
  - 150文の場合の信頼性を検証
  - 正規化のためにシステム間の相対評価値に換算して比較

	Pearson相関係数
日英	0.940
英日	0.985
中英	0.963



150文での比較結果：
多少のスコアの違いはあるが，上位のシステムと下位のシステムが入れ替わるような大きな違いはない

↓

差が大きい場合，150文での比較結果の信頼性は高い．（300文の場合はそれ以上の信頼性がある）

Adequacyの課題

- 基準が曖昧
  - 各グレードの評価が実際にどのような訳質であるか明確でない
  - 例えば，原文の意味が理解できる文や，原文の意味が理解できて文法が正しい文の割合・・・不明
- 湧き出し語に対する定義がない
- Fluencyは別評価
  - AdequacyとFluencyの役割分担が明確でない

Fluencyの課題

- Fluencyは流暢さを評価するため，原文の意味を反映していない流暢な訳文に高いスコアを与える.

  → 原文の意味を反映している訳文のみに流暢さのスコアを与えなければ，訳質の評価にならない.

＜例＞
入力：今日は晴れです.
出力：Hello!

Adequacy = 1
Fluency = 5

Acceptabilityの設計

- 各評価グレードに意味を持たせる

  ＜例＞

  - 原文の意味が理解できる
  - 原文の意味が理解できて文法が正しい

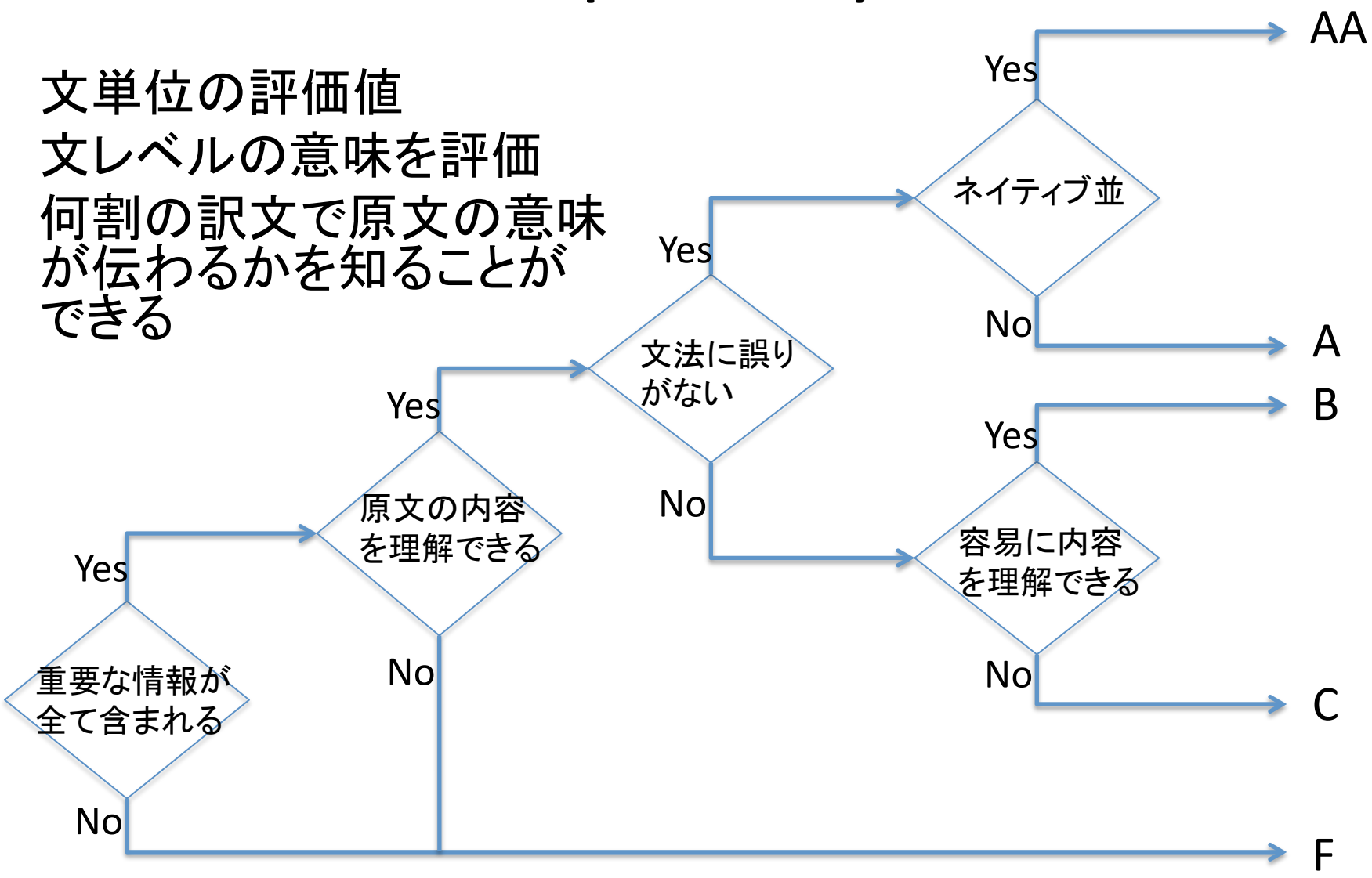
- 流暢さも同時に評価

  – 原文の意味が理解できる場合に限り，流暢さも評価

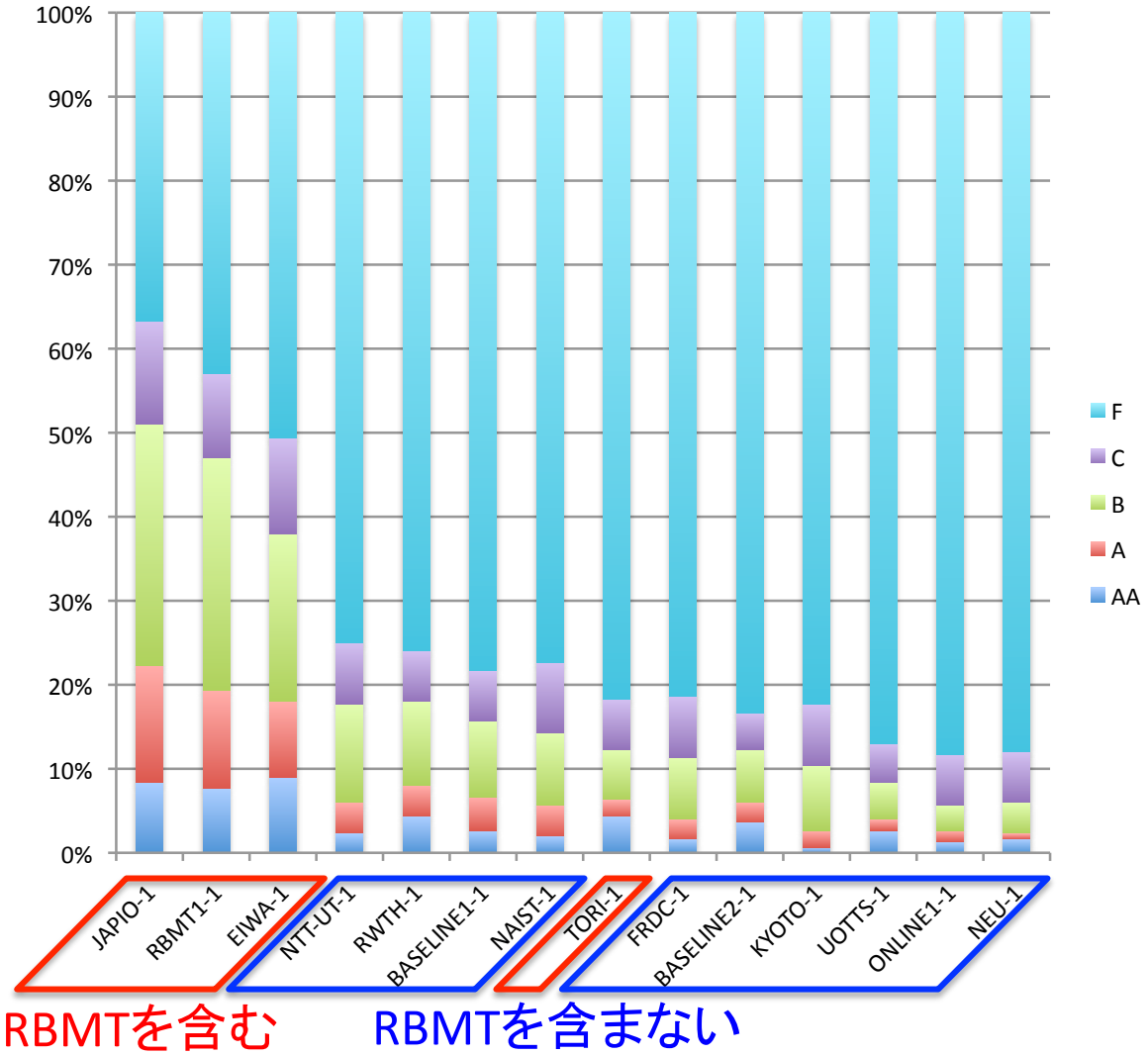
  - 原文の内容を反映しているかどうかを無視した評価を避けられる

Acceptability

- 文単位の評価値
- 文レベルの意味を評価
- 何割の訳文で原文の意味が伝わるかを知ることができる



Acceptability評価の例（NTCIR-9 日英）



- RBMTのトップシステム
  - 63%の文が理解可能
- SMTのトップシステム
  - 25%の文が理解可能

Acceptabilityの課題

- 中／低品質な訳文に対する評価の解像度が低い
  - 一部誤訳でも全部誤訳でも最低評価
  - 現状の機械翻訳の精度では，システム間の比較はAdequacyの方が向いている
- 実際の情報収集の用途での有用性とのギャップ
  - 部分的にしか正しく訳せていない文でも，重要な部分が分かれば有用である場合がある.
  - なぜこの文を翻訳する必要があるのかという目的を明確にしなければ，部分的に訳せている文の有用性を評価することは困難である.

NTCIR-10での新しい評価（特許審査評価）

- 特許審査でのMTの利用を想定した評価を実施予定
- 利用場面
  - 特許審査において，審査官が機械翻訳を利用
  - 外国語の引用文献の特許を機械翻訳して，翻訳結果から内容を理解する.
- 前提
  - 引用文献には，実際に拒絶の根拠として用いられた特許を用いる.
- 評価
  - 審査で拒絶の根拠として用いられた引用文献に記載されている事実が，翻訳結果からどれだけ認定できるか

特許審査評価のコンセプト

引用文献を機械翻訳した結果が，どの程度特許審査に有用であるか？

機械翻訳

日本語/
中国語

英語

英語

審査対象の特許文書を拒絶する根拠となる引用文献の特許文書

評価者（審査官経験者）

審査対象の
特許文書

特許審査評価の実際の枠組み

引用文献を機械翻訳した結果が，どの程度特許審査に有用であるか？

機械翻訳

評価

日本語/中国語

英語

有用性の評価は，審査で重要な引用文献中の事実を機械翻訳結果から認定できるかどうかに基づいて行う．

評価者
（審査官経験者）

準備

中国語へ人手翻訳

日本語

記載内容:
•引用文献からどのような事実が認定されたか
•なぜ特許文書Aを拒絶したか

審決（特許審査の最終決定）

特許文書Aを拒絶する根拠となった引用文献の特許文書

拒絶された特許文書A
（審査対象の特許文書）

審決の構成の一例

- 結論（特許成立・不成立）
- 引用文書の抜粋
- 引用文書から審査官が認定した事実の説明
- 審査対象の特許と引用文書との事実の比較
- 判断（特許成立・不成立の理由）

評価の流れ

- 準備
  - 「審決において引用文書から審査官が認定した事実」の根拠となる文を引用文書から抽出する
- 翻訳
  - 抽出した文を機械翻訳する
- 評価
  - 翻訳結果から，「審決において引用文書から審査官が認定した事実」をどれだけ理解できて，審査に有用であるかについて，「審査対象の特許文書（拒絶された特許）と引用文書との事実の比較」および「判断」において重要な事実かどうかを考慮して評価する

審決中での引用文書から認定した事実の例

これらの記載事項によると、引用例には、

「内部において、先端側に良熱伝導金属部43が入り込んでいる中心電極4と、
中心電極4の先端部に溶接されている貴金属チップ45と、
中心電極4を電極先端部41が碍子先端部31から突出するように挿嵌保持する絶縁碍子3と、
絶縁碍子3を挿嵌保持する取付金具2と、
中心電極4の電極先端部41との間に火花放電ギャップGを形成する接地電極11とを備えたスパークプラグにおいて、
中心電極4の直径は、1.2～2.2mmとしたスパークプラグ。」

の発明(以下「引用例記載の発明」という。)が記載されていると認められる。

認定した事実を構成要素単位に分解

1. 内部において、中心電極4の先端側に良熱伝導金属部43が入り込んでいる
2. 中心電極4の先端部に貴金属チップ45が溶接されている
3. 絶縁碍子3が中心電極4を電極先端部41が碍子先端部31から突出するように挿嵌保持する
4. 取付金具2が絶縁碍子3を挿嵌保持する
5. 接地電極11が中心電極4の電極先端部41との間に火花放電ギャップGを形成する
6. 中心電極4の直径は、1.2～2.2mm

引用文書から抽出した文の例

	認定された事実	引用文書から抽出した文（テストデータ）
1	内部において、中心電極4の先端側に良熱伝導金属部43が入り込んでいる	また、図3に示すごとく、中心電極4の内部においては、上記露出開始部431よりも先端側にも良熱伝導金属部43が入り込んでいる。
2	中心電極4の先端部に貴金属チップ45が溶接されている	また、中心電極4の先端部には、貴金属チップ45が溶接されている。
3	絶縁碍子3が中心電極4を電極先端部41が碍子先端部31から突出するように挿嵌保持する	上記中心電極4は、電極先端部41が碍子先端部31から突出するように絶縁碍子3に挿嵌保持されている。
4	取付金具2が絶縁碍子3を挿嵌保持する	上記絶縁碍子3は、碍子先端部31が突出するように取付金具2に挿嵌保持される。
5	接地電極11が中心電極4の電極先端部41との間に火花放電ギャップGを形成する	上記接地電極11は、図2に示すごとく、電極先端部41との間に火花放電ギャップGを形成する。
6	中心電極4の直径は、1.2～2.2mm	また、上記碍子固定部22の軸方向位置における中心電極4の直径は、例えば、1.2～2.2mmとすることができる。

特許審査評価　評価基準

S	審査で重要な部分の事実が全て認定できて，翻訳結果のみで審査可能
A	審査で重要である事実が半分以上認定できて，審査に有用
B	審査で重要である事実が1つ以上認定できて，審査に有用
C	一部の事実が認定できて，審査に何らかの有用性がある
D	一部の事実が認定できたが，審査に有用とはいえない
F	全く事実が認定できず，審査の役に立たない

（引用文書単位の評価）

まとめ

- 人手評価
  - 信頼性が高い評価方法
  - 時間や手間のコストは大きい → 評価するデータ量は少なくしたい
  - ランダムに選択した150文の評価でも差が大きければ信頼性は高い（ただし文数は多いほど信頼性は高い．NTCIR-9では300文を評価）

- 文の訳質の評価（NTCIR-9で実施）
  - Adequacy
    - システム比較に有用
    - 各グレードが示す訳質が明確ではない
  - Acceptability
    - 原文の意味が理解できる文数の割合が分かる

- 特許審査評価（NTCIR-10で実施予定）
  - 特許審査という応用における有用性の評価
  - 実施はこれから
  - 評価結果の発表は2013/06のNTCIR-10 Workshopにて