# 大規模特許対訳コーパスによる英日／中日／韓日統計翻訳システムの性能評価

Evaluation of Statistical Machine Translation System using Large-scaled Parallel Corpus in Patent Translation.

株式会社東芝　インダストリアル ICT ソリューション社　**園尾　聡**

2009 年九州工業大学大学院生命体工学研究科博士課程修了。博士（工学）。同年株式会社東芝入社。AAMT/Japio 特許翻訳研究会 拡大評価部会メンバー。自然言語処理の研究に従事。

satoshi.sonoo@toshiba.co.jp

## 1 はじめに

外国語で書かれた大量の特許文書を検索・調査するために機械翻訳技術が広く用いられている。機械翻訳技術には、大きく二つの枠組みがある。一つは、長年研究開発が続けられてきた、辞書や訳し分け知識を記述した規則に基づく規則ベース機械翻訳である。もう一つは、対訳コーパスを利用した統計的機械翻訳（Statistical Machine Translation; 以下 SMT）[1] であり、近年盛んに研究が行われている。SMT の一手法であるフレーズベース SMT では、対訳コーパスから対訳となるフレーズを統計的に抽出し、翻訳モデルを構築し、翻訳を行う。

特許翻訳においては、外国出願されているファミリー特許を利用して大規模な対訳コーパスが整備され、数百万規模の対訳コーパスを用いた特許翻訳ワークショップも開催されている [2,3]。また、みんなの自動翻訳 @ TexTra® [4] では、3000 万文の対訳コーパスを用いた英日 SMT エンジンが利用されている。さらに 2015 年には、JPO および NICT から数千万～数億規模の英日、中日、韓日対訳特許コーパス [5] が公開された。これほど大規模な対訳コーパスの公開の例はなく、特許翻訳においてコーパスサイズに対する SMT の翻訳性能への効果は明からにされていない。

本研究では、JPO および NICT から公開された大規模な対訳コーパスを用いて英日・中日・韓日の SMT エンジンを構築し、性能評価を行ったので、その内容について報告する。

## 2 特許コーパスの概要

本研究では、JPO·NICT 英日対訳コーパス 3.5 億文、JPO 中日対訳コーパス 1.3 億文、JPO・NICT 韓日対訳コーパス 0.8 億文を利用した。各コーパスとも、化学、電気、機械、物理の 4 分野から構成される。各コーパスの分野ごとの対訳文数を表 1 に示す。また、それぞれの対訳文には、対訳文としての整合性を示すアライメントスコアおよび、文対応情報（日本語 M 文に対して、対象言語 N 文が対応）が付与されている。SMT では、一般的にコーパスサイズを増やすことにより翻訳精度が向上するが、アライメントスコアが低い文が含まれている場合はノイズとなってしまい、一定量を超えると翻訳精度の伸びが鈍化または低下する。したがって、コーパス全体からアライメントスコアが高い順に対訳文を選択し、訓練コーパスとして用いた。さらに、中日・韓日方向については、文対応情報が 1 対 1 の対訳文に限定して利用した。

表 1　特許コーパスの文数

| 分野 | 文数（百万文） | | |
|---|---|---|---|
| | 英日 | 中日 | 韓日 |
| 化学 | 128.1 | 50.7 | 33.7 |
| 電気 | 86.5 | 24.7 | 23.5 |
| 機械 | 45.3 | 16.3 | 8.2 |
| 物理 | 88.0 | 31.1 | 18.1 |
| 計 | 348.0 | 132.9 | 83.5 |

## 3 SMT エンジンの構築

SMT エンジンの構築は、オープンソースコードである Moses[1] を用いて行った。言語モデルは、5-gram、アライメントモデルは、" grow-diag-final"、リオーダリングモデルは、" wbe-msd-bidirectional-fe"、翻訳モデルは、大規模コーパスに対しても比較的訓練時間が短くて済むフレーズベースモデルを採用した。

単語分割に関しては、日本語は、Mecab[2]、英語は、Moses に含まれる tokenizer.perl、中国語は、Stanford Word Segmenter[3]、韓国語は、Mecab-ko[4] を利用した。また、distortion_limit に関しては、英日·中日方向は 20、韓日は、言語類似性を考慮し、0 に設定した。重み最適化は、k-best batch MIRA[6] によって行い、重み最適化および自動評価には、英日は NTCIR-10 Patent MT[2]、中日・韓日は WAT2015 Patent Subtasks[3] にそれぞれ含まれる開発・テストセットを用いた。

## 4 英日 SMT エンジンの性能評価

英日 SMT エンジンにおける、コーパスサイズに対する自動評価スコア(BLEU)の変化を図 1 に示す。ここで、点は実験値を示し、実線は近似曲線である。図 1 では、2000 万文までは翻訳精度の顕著な向上が見られ、さらにアライメントスコアが 0.2 以上に相当する 1.2 億文まで増加させた場合においても翻訳精度の一定の向上が確認された。

一方、これ以上のコーパスサイズを直接扱うことは計算機リソースの制約上困難であったため、分野別の訓練を試みた。すなわち、コーパス全体を分野別に分割し、分野別コーパスで訓練された翻訳モデルを対数線形結合することで、全てのコーパスを用いた翻訳エンジンとして評価した。その結果、1.2 億文を用いた際の BLEU 値（単語単位）が 42.40 に対して、全てのコーパス[5] を用いた場合は 61.01 となり、さらに高い翻訳精度が得られることを確認した。

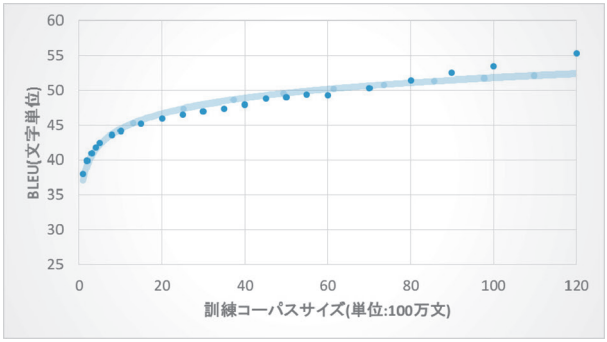



図 1　英日統計翻訳エンジンにおける自動評価（BLEU）

次に、コーパスと共に配布された品質評価用シートに対して、特許庁「特許文献機械翻訳の品質評価手順」[7] に従って人手評価を実施した。評価基準は以下の通りである：

「内容の伝達レベル」

5：すべての重要情報が正確に伝達されている。（100%）
4：ほとんどの重要情報は正確に伝達されている。(80%～)
3：半分以上の重要情報は正確に伝達されている。(50%～)
2：いくつかの重要情報は正確に伝達されている。(20%～)
1：文意がわからない、もしくは正確に伝達されている重要情報はほとんどない。（～ 20%）

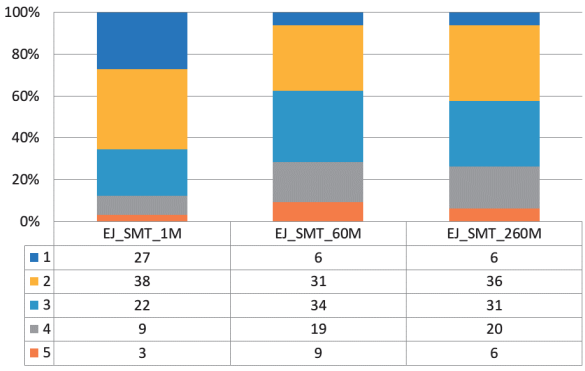
「重要技術用語の翻訳精度」

A（適訳語）：人手翻訳に照らし、技術的に同義かつ一般的に用いられる訳語である。
B（可訳語）：技術用語として一般的に用いられる訳語ではないが、意味はおおむね正しい。
C（誤訳語）：誤訳である。
D（不訳語）：未知語、訳漏れである。

評価対象は、100 万文のコーパスを用いた SMT エンジン（EJ_SMT_1M）、同 6000 万文（EJ_SMT_60M）、同 2.6 億文（EJ_SMT_260M）である。人手評価の結果を図 2 および図 3 に示す。内容の伝達レベルの観点（図 2）では、内容がほぼ理解可能な翻訳結果（伝達レベル 4 以上）の割合が、EJ_SMT_1M では 12.0% に対し、EJ_SMT_60M では 28.0%、EJ_SMT_260M では 26.0%、と人手評価において

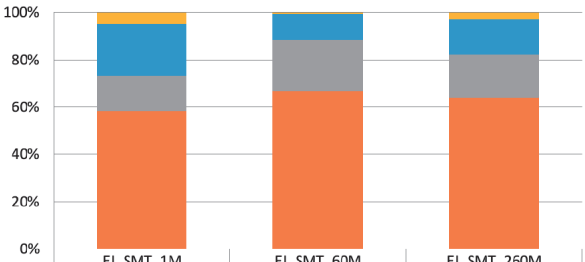
1　http://www.statmt.org/moses
2　http://taku910.github.io/mecab/
3　http://nlp.stanford.edu/software/segmenter.shtml
4　https://bitbucket.org/eunjeon/mecab-ko/
5　実際には、文対応が 1 対 1 以外、アライメントスコアが 0、明細書以外を除いた約 2.6 億文を用いた。

もコーパスサイズに応じて翻訳精度が向上されることが確認された。しかしながら、EJ_SMT_60M と EJ_SMT_260M を比べても、自動評価で見られたような顕著な向上は確認されなかった。重要技術用語の翻訳精度の観点（図 3）においては、コーパスサイズを増やすことによって、より専門的な技術用語が訳出される傾向が見られたが、全体的にはフレーズテーブルの曖昧性が増えてしまい、EJ_SMT_260M では多少悪化するという結果となった。




| | EJ_SMT_1M | EJ_SMT_60M | EJ_SMT_260M |
|---|---|---|---|
| 1 | 27 | 6 | 6 |
| 2 | 38 | 31 | 36 |
| 3 | 22 | 34 | 31 |
| 4 | 9 | 19 | 20 |
| 5 | 3 | 9 | 6 |

図 2　英日統計翻訳エンジンにおける人手評価（内容の伝達レベル）




| | EJ_SMT_1M | EJ_SMT_60M | EJ_SMT_260M |
|---|---|---|---|
| D | 9 | 1 | 5 |
| C | 39 | 20 | 27 |
| B | 27 | 39 | 33 |
| A | 105 | 120 | 115 |

図 3　英日統計翻訳エンジンにおける人手評価（重要技術用語の翻訳精度）

## 5 中日 SMT エンジンの性能評価

中日 SMT エンジンにおける、コーパスサイズに対する自動評価スコアの変化を図 4 に示す。英日方向と同様に、コーパスサイズの増加に伴って翻訳精度が向上していることが確認できた。特に 2000 万文あたりまで大幅な向上が見られ、その後、割合は小さいものの全てのコーパス（1.3 億文）を用いた場合まで一定の向上傾向が見られた。

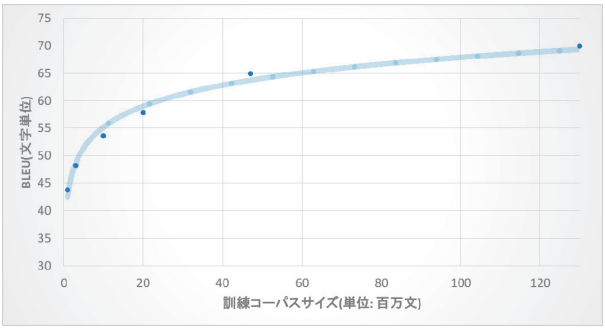
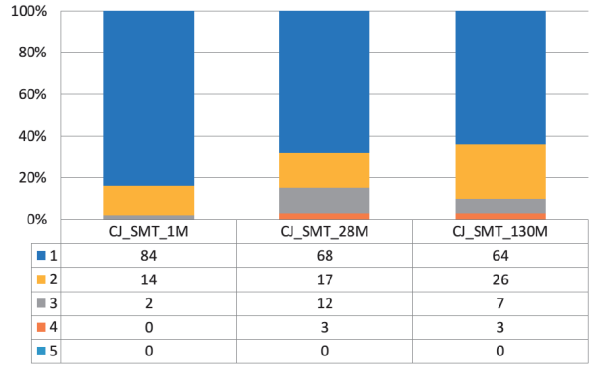


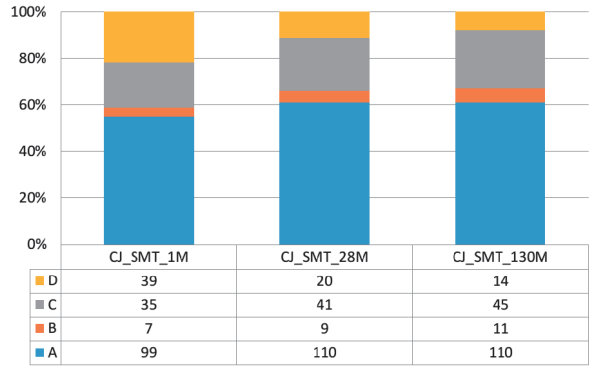

図 4　中日統計翻訳エンジンにおける自動評価（BLEU）

英日方向と同様に人手評価を実施した。100 万文のコーパスを用いた SMT エンジン（CJ_SMT_1M）、同 2800 万文（CJ_SMT_28M）、同 1.3 億文（CJ_SMT_130M）に対する評価結果を図 5 および図 6 に示す。内容の伝達レベルの観点では、全てのコーパスを用いた場合でも内容の理解度は、高くない。その中で、CJ_SMT_1M に比べて CJ_SMT_28M および CJ_SMT_130M では、重要情報がある程度以上伝達される翻訳結果（伝達レベル 2 以上）の割合が改善している。一方で、内容の伝達レベルおよび重要技術用語の翻訳精度において、CJ_SMT_28M と CJ_SMT_130M の間に有意な差は見られなかった。




| | CJ_SMT_1M | CJ_SMT_28M | CJ_SMT_130M |
|---|---|---|---|
| 1 | 84 | 68 | 64 |
| 2 | 14 | 17 | 26 |
| 3 | 2 | 12 | 7 |
| 4 | 0 | 3 | 3 |
| 5 | 0 | 0 | 0 |

図 5　中日統計翻訳エンジンにおける人手評価（内容の伝達レベル）




| | CJ_SMT_1M | CJ_SMT_28M | CJ_SMT_130M |
|---|---|---|---|
| D | 39 | 20 | 14 |
| C | 35 | 41 | 45 |
| B | 7 | 9 | 11 |
| A | 99 | 110 | 110 |

図 6　中日統計翻訳エンジンにおける人手評価（重要技術用語の翻訳精度）

## 6 韓日 SMT エンジンの性能評価

韓日 SMT エンジンにおける、コーパスサイズに対する自動評価スコアの変化を図 7 に示す。英日・中日方向とは異なり、韓日方向では、2600 万文（アライメントスコアが 0.2 以上に相当）辺りに翻訳精度のピークがあり、それ以上コーパスを増やしても翻訳精度は低下傾向となることが確認された。BLEU 値は総じて非常に高い値（70 以上）を示しており、もともと翻訳性能が高い韓日 SMT エンジンにおいて、コーパスサイズを増加させることにより開発セットに過剰調整された可能性がある。

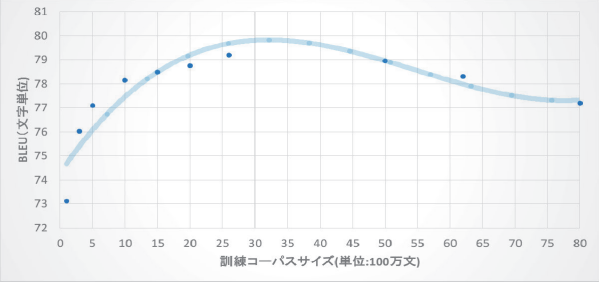
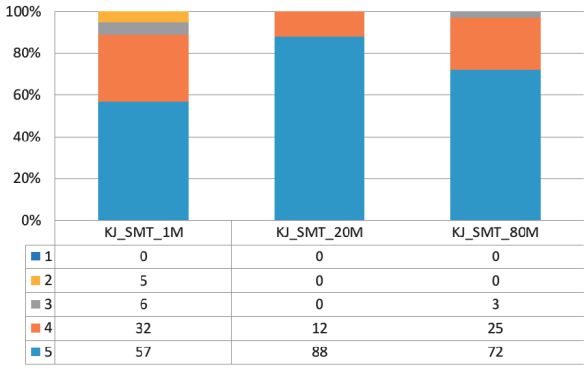


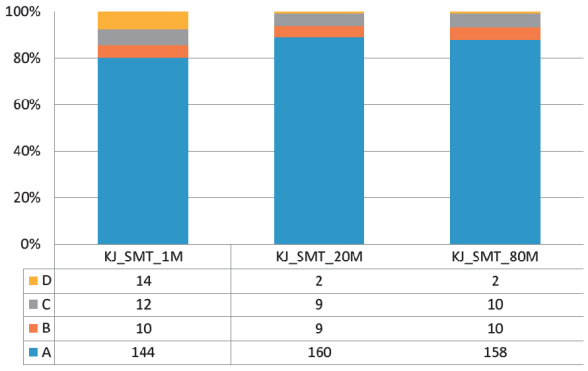

図 7　韓日統計翻訳エンジンにおける自動評価（BLEU）

英日・中日方向と同様に人手評価を実施した。100 万文のコーパスを用いた SMT エンジン（KJ_SMT_1M）、同 2000 万文（KJ_SMT_20M）、同 8000 万文（KJ_SMT_80M）に対する評価結果を図 8 および図 9 に示す。内容の伝達レベルの観点では、各 SMT エンジンにおいて、内容がほぼ理解可能な翻訳結果（伝達レベル 4 以上）の割合が 80% を超える翻訳精度を示した。その中で、KJ_SMT_20M が最も高い内容伝達レベルを示した。KJ_SMT_80M では不要な湧き出し語や訳抜けによる精度低下が生じ、自動評価結果とも整合する結果が得られた。重要技術用語の翻訳精度に関しては、KJ_SMT_20M および KJ_SMT_80M においてほぼ同等の評価結果が得られた。



| | KJ_SMT_1M | KJ_SMT_20M | KJ_SMT_80M |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 5 | 0 | 0 |
| 3 | 6 | 0 | 3 |
| 4 | 32 | 12 | 25 |
| 5 | 57 | 88 | 72 |

図 8　韓日統計翻訳エンジンにおける人手評価（内容の伝達レベル）



| | KJ_SMT_1M | KJ_SMT_20M | KJ_SMT_80M |
|---|---|---|---|
| D | 14 | 2 | 2 |
| C | 12 | 9 | 10 |
| B | 10 | 9 | 10 |
| A | 144 | 160 | 158 |

図 9　韓日統計翻訳エンジンにおける人手評価（重要技術用語の翻訳精度）

## 7 おわりに

本研究では、数千万文～数億文規模の特許対訳コーパスを用いて英日・中日・韓日 SMT エンジンの性能評価を行った。英日・中日方向では、全てのコーパスを利用することで、自動評価（BLEU）による精度向上が見られたが、それに対応するような人手評価結果は確認できなかった。ただし、これは人手評価文が不足していることにも起因しており、より詳細な評価が必要である。韓日方向では、全てのコーパスを使うよりもアライメントスコアの高い一部のコーパスを利用する方が、自動評価および人手評価において高い評価結果を得られた。

一方、大規模な特許対訳コーパスを単純に利用するだけでは、SMT エンジンの訓練および運用に莫大な計算リソースが必要となる。特許分類が明確な特許文書の特性を踏まえ、大規模な特許対訳コーパスが利用可能であれば、統計翻訳モデルを分野別（特許分類別）に構築することで、計算機リソースを抑え、高精度な特許翻訳システムの実用化が期待できる。

## 謝辞

本研究は、高度言語情報融合フォーラム（ALAGIN）が提供する「JPO・NICT 英日対訳コーパス」、「JPO・NICT 韓日対訳コーパス」、「JPO 中日対訳コーパス」のコーパスを利用した成果である。コーパスの公開および研究利用にご協力頂いた関係各位に対し、謝意を表します。

## Appendix：

各 SMT エンジンの人手評価においてコーパスサイズによる改善効果が見られた翻訳結果の一例を表 2 に示す。

## 参考文献

[1] Koehn, P.（2009）. Statistical machine translation. Cambridge University Press.

[2] Goto, I., Chow, K. P., Lu, B., Sumita, E., & Tsou, B. K.（2013）. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In NTCIR.

[3] Nakazawa, T., Mino, H., Goto, I., Neubig, G., Kurohashi, S., & Sumita, E.（2015）. Overview of the 2nd Workshop on Asian Translation. In Proc. of the 2nd Workshop on Asian Translation（WAT2015）.

[4] 内山 将夫（2014）、英日・中日特許 SMT システムの実用化と課題、Japio YEAR BOOK 2014, pp.252-255.

[5] ALAGIN 言語資源・音声資源サイト , https://alaginrc.nict.go.jp/

[6] Cherry, C., & Foster, G.（2012）. Batch tuning strategies for statistical machine translation. In Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies（pp. 427-436）.

[7] 特許文献機械翻訳の品質評価手順について , https://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm

表２　各 SMT エンジンにおける翻訳結果の一例

SRC：原文

REF：対訳コーパスにある訳文

EJ_SMT_＊M、CJ_SMT_＊M、KJ_SMT_＊M：SMT エンジンの翻訳結果

| | |
|---|---|
| SRC | A directional drilling machine 1 then drills rows of pilot holes under the site, which define the basin's elongated shape. |
| REF | 次いで**傾斜掘削機**1が、盆地状の細長い形状を規定する一連の案内孔を用地の下に穿孔する。 |
| EJ_SMT_1M | 指向性穿孔機1の下に細長い形状の容器を定義するサイトの行ドリルパイロット孔。 |
| EJ_SMT_260M | 次に、指向性ドリル装置1がサイトの下に案内孔を列状に穿孔して、水盤の細長い形状を画定する。 |
| SRC | Sea water, however, contains significant quantities of dissolved oxygen, about 10 ppm at 100 C, which renders it unsuitable for use in its raw state because of its corrosive action and its encouragement of bacterial growth. |
| REF | しかし海水は、相当な量、すなわち、10 ℃で約10 ppmの溶融酸素を含んでおり、それはその**腐食作用**およびバクテリア増殖促進作用のために、海水をそのままの状態で使用することに適さない。 |
| EJ_SMT_1M | しかし、かなりの量の溶解酸素を含む、約10ppm海水で100での使用には不適であるので、これはその腐食性動作およびそのencouragement生状態Cの細菌成長の。 |
| EJ_SMT_260M | しかしながら、海水は、100℃で、約10ppm、溶存酸素を相当量含有し、その生の状態で使用するのに不適切であり、その腐食作用のために、細菌増殖を助長する。 |
| SRC | 故本研究将黄原胶和瓜儿豆胶进行复配，将凝胶剂总用量、黄原胶与瓜儿豆胶的比例以及溶胶温度这三个对本凝胶性能影响最大的因素作为考察对象，采用凝胶强度测定仪分别测得不同条件所得凝胶的强度。 |
| REF | そのため、本研究ではキサンタンガムとグアーガムを配合し、ゲル製剤の総使用量、キサンタンガムとグアーガムの割合及びゲルの溶解温度という3つの本ゲル製剤の性能に対して最大の影響を及ぼす要素を考察の対象とし、ゲル強度測定装置を採用して異なる条件の下で得られるゲルの強度をそれぞれ測定した。 |
| CJ_SMT_1M | すなわち、研究黄原胶と瓜儿豆胶を复配ゲルは、総用量、黄原胶と瓜儿豆胶の割合と、ゾル温度の三つの本ゲル性能の要因によって最も影響を考察する対象として、ゲル強度測定器を用いて測定した得られたゲルは、それぞれ異なる条件の強度を示している。 |
| CJ_SMT_130M | 従って、この研究は、キサンタンガムおよびグアーガム、キサンタンガム、グアーガム、およびソルゲル複合を行い、総使用量の割合に依存して、ゲル強度試験器によるゲル特性ゲル特性に最も影響を与える因子を考察の対象として温度の3つの異なる条件、得られたゲルの強度をそれぞれ測定した。 |
| SRC | 水解过程采用一种以上蛋白酶时，可根据不同蛋白酶适宜水解的PH值和温度，选择采用同时加入或依次加入进行水解。 |
| REF | 加水分解プロセスにおいて、一種類以上のプロテアーゼを使用する場合、異なるプロテアーゼの加水分解に適したPH値と温度に基づいて、同時に添加または順番に添加して加水分解を行うことを選択できる。 |
| CJ_SMT_1M | 加水分解のプロセスは、一種類以上プロテアーゼの際には、異なる適宜プロテアーゼ加水分解のpHと温度を採用するかを選択すると共に、を添加して加水分解を行った。 |
| CJ_SMT_130M | 加水分解の間、加水分解に適切なpHおよび温度に依存して、1つ以上のプロテアーゼを用いてプロテアーゼを模す場合には、同時にまたは連続的に添加され、加水分解した。 |
| SRC | 본 발명은 유상 부, 수상 부 및 영양성분 부를 포함하는 화장료에 있어서, 대두에서 추출한 인지질 및 에몰리언트제가 함유된 유상 부를 정제수, 글리세린 및 1,3-부틸렌글리콜이 함유된 수상 부에 1차 유화하여 리포좀을 형성시킨 다음, 여기에 생리활성물질이 함유된 영양성분 부를 캡슐화시켜 2차 유화하고, 점증제 부와 기타 첨가제 부를 첨가하여 제조된 나노사이즈의 인지질 리포좀 화장료 및 그 제조방법을 그 특징으로 한다. |
| REF | 本発明は、油相部、水相部及び栄養成分部を含む化粧料において、大豆から抽出したリン脂質及びエモリエント剤が含まれた油相部を、精製水、グリセリン及び1,3-ブチレングリコールが含まれた水相部に1次乳化してリポソームを形成させた後、ここに生理活性物質が含まれた栄養成分部をカプセル化させて2次乳化し、増粘剤部とその他添加剤部を添加して製造されたナノサイズのリン脂質リポソーム化粧料及びその製造方法をその特徴とする。 |
| KJ_SMT_1M | 本発明は、有償部、水相部および栄養成分部を含む化粧料において、大豆から抽出した燐脂質及びにUNK몰리アンドット剤が含まれた有償部を精製数、グリセリン及び1、3−부틸렌グリコールが含有された水相部に1次乳化してリポソームを形成した後で，ここに生理活性物質が含まれた栄養成分部をカプセル化させ、2次乳化し、漸増制御部とその他添加剤部を添加して製造されたナノサイズの燐脂質リポソーム化粧料及びその製造方法をその特徴とする。 |
| KJ_SMT_20M | 本発明は、油相部、受像部および栄養成分部を含む化粧料に大豆から抽出したリン脂質及びエモリエント剤が含有された油相部を精製水、グリセリンと1，3−ブチレングリコールが含有された水相部に1次乳化してリポソームを形成させた後、ここに生理活性物質が含有された栄養成分部をカプセル化させ、2次乳化し、増粘剤及びその他の添加剤を添加して製造されたナノサイズのリン脂質リポソーム化粧料及びその製造方法をその特徴とする。 |
| SRC | 전기가열에 의한 점화기와 반응물을 연소/개질하는 촉매를 구비하는 탄화수소 연소용 촉매반응기에 있어서, 반응물이 도입되어 연소가 개시되는 착화부와 상기 착화부를 지난 반응물이 개질되는 주반응부를 포함하고, 상기 착화부의 단면적이 상기 주반응부의 단면적보다 작게 형성되며,상기 점화기는 상기 착화부에 설치되는 것을 특징으로 하는 탄화수소 연소용 촉매반응기. |
| REF | 電気加熱による点火器と反応物を燃焼/改質する触媒を備えた炭化水素の燃焼用触媒反応器において、反応物が導入されて燃焼が開始される着火部と、上記着火部を通過した反応物が改質する主反応部を含み、上記着火部の断面積が上記主反応部の断面積よりも小さく形成され、上記点火器は、上記着火部に設置されることを特徴とする、炭化水素燃焼用触媒反応器。 |
| KJ_SMT_1M | 電気加熱によるUNK점화기と反応物を燃焼／改質する触媒を具備する炭化水素燃焼用触媒反応器において、反応物が導入されて燃焼が開始される着火部と上記着火部を過ぎた反応物が改質される주반部を含み、上記着火部の断面積が上記UNK주반部の断面積より小さく形成され、前記점화기は上記着火部に設置されることを特徴とする炭化水素燃焼用触媒反応器。 |
| KJ_SMT_20M | 電気加熱による点火部と反応物を燃焼・改質する触媒を備える炭化水素燃焼用触媒反応器において、反応物が導入されて燃焼が開始される着火部と前記着火部を過ぎた反応物が改質される主反応部を含み、前記着火部の断面積が前記主反応部の断面積より小さく形成され、前記イグナイタは、上記着火部に設けられることを特徴とする炭化水素燃焼用触媒反応器。 |