異言語資源を利用したモチーフラベルの自動推定

三澤賢祐　　松本裕治
奈良先端科学技術大学院大学

{kensuke-mi, matsu}@is.naist.jp

1　はじめに

現在，我々は「ペルシア語口承文芸コーパス (*Iranian Folk Narratives*:IFN)」[9] のデータ整備を行なっている．その作業の 1 つとして，物語へのラベル付与がある．民族学・文化人類学の分野では古くから物語の研究が行われており，個々の物語の類似点と相違点を調べるため，物語へのラベル付与が考案されてきた．付与されたラベルを比較，検討することで，文化的な発見，または物語類型の分析が可能になる．現在，一般的に用いられているラベルは「トンプソンのモチーフインデックス分類法」(*Thompson Motif Index*:TMI)[8] である．これは

- 世界中の民話・神話のモチーフをカバーしている
- 階層的に設計されており，効率的な分類が可能

といった利点から物語の分類には頻繁に用いられている [1][10]．TMI には，例えば *F51 Sky-rope. Access to upper world by means of a rope.* というラベルがある．イギリスの有名な童話「ジャックと豆の木」の空に登るシーンや，熊本県の昔話「天道さん金ん網」[11] など，世界中の物語の，天空に登るシーンが「*F51*」として記号化されている．TMI のラベル総数は約 5 万あり，それらは階層的に構成されており，その最上位階層は A-Z から成る 23[1]のラベルにまとめられている．図 1 に，A ラベル:*Mythological Motifs* の一部を示した．TMI は物語中の任意の場所にラベルを付与でき，物語文書自体，または文書中の文を対象にラベルを与えることも可能である．また，1 つの物語文書中に複数のラベルを付与することができる．

しかしながら，TMI の，人手によるラベル付与はコストが高い．そのため，近年，このコストを下げるために，教師あり学習を用いた TMI ラベル推定モデルが提案された．Karsdorp ら [2] はオランダ語民話データベース (*Dutch Folktale Database*:DFD)[7] で，TMI の自動ラベル付与を行なった．TMI ラベルは，1 つの物語文書に複数のラベルが付与できるため，自動 TMI ラベル付与はマルチラベル問題として定式化できる．Karsdorp らはこの問題を解くために，Okapi-BM25 とラベル付き LDA(L-LDA) を用いた．DFD を訓練データとテストデータに分割し，評価を行なった．その結果，Okapi-BM25 では 78%，L-LDA では 72%の Average Precision を達成している．

本研究では IFN への自動ラベル付与を目的としている．しかし，IFN のうちラベル付けがされた物語文書はわずか 5%に過ぎず，Karsdorp らのように同一ドメインだけで精度の高いモデルの構築は困難である．また，ラベル付けされた物語文書の電子データは，我々の知る限り DFD 以外に大規模には存在していない．そこで，我々は DFD を訓練データとして用いる．ただし，DFD はオランダ語で記述されたコーパスであるため，直接は IFN へのラベル付与の訓練データとして用いることはできない．そこで提案手法では，機械翻訳によってオランダ語とペルシア語を一旦英語へと翻訳し，モデルの構築を実施する．

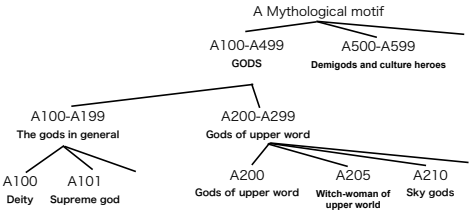

図 1: モチーフインデックスラベルの木構造の例

表 1: DFD の統計

	All	TMI ラベル付きのみ
文書数	42,885	1,452
文数	568,298	31,792
述べ語数	8,251,950	664,584
異なり語数	234,520	35,421

2　IFN へ自動ラベル付けをするために使用できる言語資源

2.1　*Dutch Folktale Database*:DFD

DFD[7] は，オランダの *Meertens Institute* により開発，公開されている物語データベースである．データベース中の物語はオランダ語，もしくはフリスク語で記述されており，様々なメタデータが利用可能である．TMI もオランダ語で記述された 1,452 話にアノテーションされている．なお，DFD では物語文書自体に TMI ラベルが付与されている．表 1 に DFD の中で TMI を利用可能な物語文書の統計情報を示す．

2.2　*Thompson Motif Index* 説明文:TMI 説明文

TMI は Web 上で公開されており[2]，以下の情報で構成されている．ラベル説明文も訓練データとして用いた．

モチーフ番号:A101
モチーフ番号の説明文:Supreme god as creator
地域と文献名:Armenian,Ananikian 20

[1] I と O と Y は欠番ラベルである．

[2] http://www.ruthenia.ru/folklore/thompson/

表 2: IFN のデータ

	All	TMI ラベル付きのみ
文書数	325	16
文数	6,129	806
述べ語数	91,193	10,285
異なり語数	8,807	1,509

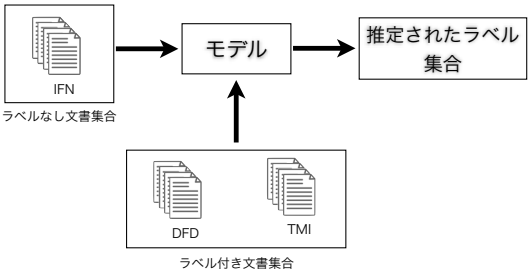

図 2: 翻訳済みの DFD と TMI を訓練データとしてモデルを構築する

2.3 *Iranian Folk Narratives*:IFN

IFN とは，イラン地域の口承（口伝えのみで伝承されている物語）文学を記録したコーパスである [9]．コーパス中の物語は口語ペルシア語で記述されており[3]，一部の物語ではモチーフインデックスもメタデータとして利用可能である．今回はこれを評価に利用した．表 2 に IFN の統計情報を示す．

3 機械翻訳を用いたモチーフインデックスラベル自動推定の提案手法

機械翻訳による言語統一

IFN 全体に TMI を自動付与しようと思ったとき，現状では十分な訓練データを用意できない．しかし，オランダ語ならば DFD で大量のラベル付き文書を用意できる．そこで，オランダ語で記述された文書を翻訳し，訓練データとして利用する．本研究では，事前に中間言語を英語に設定し，翻訳する．中間言語として英語を選んだ理由は，言語族の近さと翻訳資源の豊富さのためである．また，今回は TMI ラベルを付与する対象は文書とする．

3.1 Okapi-BM25 を用いた文書類似度による TMI ラベル推定手法

Karsdorp らは TMI 自動付与をランキング問題として定式化し，Okapi-BM25[6] を利用した．ラベル付きの訓練コーパスから，同じラベルを持つ文書のみを集め，疑似文書を生成する．ラベルが未知の文書に対してラベルを推定する時には，生成した疑似文書との類似度スコア S を計算し，類似度スコア S が高い順にランキングを行う．

Okapi-BM25 は次の式で定義される．

$$S(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{avgdl})}$$

[3]Pilevar[5] によると，口語ペルシア語と文語ペルシア語は形態素レベルと統語レベルで異なっており，現在のところ口語ペルシア語を文語ペルシア語に変換する手法は存在していない

D は文書集合，Q は単語集合で，f は頻度を数える関数である．IDF は次の式で求められる．

$$IDF(q_i) = log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}$$

また，N はコーパス中に存在する文書数，$n(q_i)$ は単語 q_i を含む文書数である．b と k_1 は任意に設定されるパラメータで，本研究ではグリッド探索を行い，最適な値にチューニングした．

Karsdorp らは Okapi-BM25 で類似度ランキングを行なったが，表層単語だけを用いる場合，作成した疑似文書中の単語と完全に一致する単語でないと，類似度計算の際に類似度スコアスコア S に加算されない．そこで，WordNet[3] を用い，単語の同義語 (*synset*) と上位概念語 (*hypernym*) も考慮に入れた文書類似度によるランキングを行う．同義語と上位概念語の追加により，類似した単語も無視されずに類似度の計算が可能である．具体的には，生成した疑似文書に含まれる全単語の同義語と上位概念語を WordNet により得て，疑似文書に追加する．

3.2 One-Versus-Rest 分類器による TMI ラベル自動推定

マルチラベル問題の解き方の一つに *Binary Relevance(BR)* と呼ばれる問題変形がある．これは，複数存在するラベルごとに二値分類器を作成し，個々の分類器が独立にラベルの有無を判定する手法である．この分類器には One-Versus-Rest 分類器が用いられることが多い．

そこで，提案手法でも One-Versus-Rest 分類器による TMI ラベル推定を行う．素性には *bag-of-words* 素性と TFIDF 値を元にした素性を利用し，その性能を比較した．

また分類に有効な単語のみを素性として用いるために，単語選択を行なった．具体的には，L2 ノルムで正規化を行なった TFIDF 値に閾値を設け，閾値以上の TFIDF 値を持つ語のみを素性として用いる．閾値の設け方は次の 2 通りとした．

- 正規化済み TFIDF 値から平均値を算出し，この平均値を閾値とする．この素性選択の方法を表 4 では TFIDF1 と表記する．
- 平均値を求める際に，ラベルごとに平均値を求め，閾値とする．表 4 では TFIDF2 と表記する．

4 実験設定

DFD を 19:1 の割合で訓練用とテスト用に分割し，訓練用に 1,437 話，評価用に 15 話を用意した．DFD を英語に翻訳する際には，Neubig ら [4] のシステムを利用した．

IFN は Google 翻訳（2013 年 11 月 10 日）用いて英語へと翻訳をした．評価には TMI ラベル付きの 16 話を利用する．なお，口語ペルシア語と文語ペルシア語は形態素レベルと統語レベルで異なっており，口語ペルシア語のままでは精度の高い機械翻訳が望めない．そこで，人手によって作成された口語文語対応表を使い，文語ペルシア語に変換した後，英語に翻訳した．

TMI 説明文は英語で記述されているので，翻訳の必要はない．

訓練とテストに用いる文書は，英語に翻訳後，すべて小文字化し，空白スペースで tokenize し，単語はすべて lemma 化を行なった．lemma 化には nltk ver.2.0.4

表 3: Okapi-BM25 によるランキング結果（AP が average precision を表している）

コーパス	タイプ	k	b	1-Error	coverage	AP	ranking loss
IFN	表層単語	0.5	0.01	0.6875	5.8125	0.4394	0.75
	表層単語 +同義語 +上位概念語	0.5	0.01	0.6875	5.5	0.4457	0.75
DFD	表層単語	0.5	0.01	0.86	6.2	0.37	0.78
	表層単語 +同義語 +上位概念語	0.5	0.01	0.73	6.2	0.41	0.78
	Karsdorp	1.2	0.75	0.26	10.69	0.78	0.27

の WordNetLemmatizer モジュールを利用した．また，nltk で定義されているストップワードの除去を行なった．

線形分類器には liblinear ver.1.9.4 を用いた．パラメータチューニングには，グリッド探索で最適なパラメータ C を求めた．また正則化項には L1 ノルムと L2 ノルムの両方を利用し，その性能を比較した．

評価にはマルチラベル問題で利用される評価基準を用いた．マルチラベル問題の評価基準は大きく 2 つに分けられる．

- 二値分類問題の評価基準
- ランキング問題の評価基準

二値分類モデルの評価には，*Hamming Loss, Subset Accuracy, Precision, Recall, F_1 measure, Accuracy* を用いた．ランキングモデルの評価には，*1 Error, Coverage, Ranking Loss, Average Precision* を用いた．

Okapi-BM25 には 2 つのハイパーパラメータの k と b が存在する．そこで k と b の値をグリッド探索により変化させ，性能を評価した．k の値が 0.01-4，b の値が 0.01-1 までの範囲でグリッド探索を行なった．

5 モチーフインデックスラベル自動推定の実験結果

5.1 Okapi-BM25 によるランキングによる実験結果

グリッド探索をした結果のうち，最高数値の結果のみを表 3 に示した．また，Karsdorp[2] らが行なった DFD のインドメインでの実験結果も参考までに示す．

表層単語だけでなく，同義語と上位概念語を追加した時の方が，coverage 値が小さくなり，また Average Precision の値が上昇している．したがって，同義語と上位概念語の追加は有意であったと言える．

5.2 One-Versus-Rest 線形分類器による実験結果

One-Versus-Rest 線形分類器を使った手法の実験結果を表 4 に示す．

IFN で最も性能が高かったのは素性選択が TFIDF1 で，正則化項が L2 の時であった．

bag-of-words 素性の時には，DFD では L1，L2 ともに F1 値が 35%程度で分類できていた．しかし，IFN では L1 正則化で 9%程度，L2 正則化で 3%程度の F1 であった．IFN での性能が低かった原因は，DFD のドメインに特化した学習がされたため，IFN のドメインへの汎化が難しかったと考えられる．

TFIDF1 の素性選択では，DFD ドメインでは性能が下がったものの，IFN ドメインでは性能は向上した．これは素性選択により，分類に有効な素性が選択されたと考えられる．

TFIDF2 の素性選択では，DFD ドメインでは *bag-of-words* 素性よりも性能が向上したが，一方で IFN ドメインでは低かった．TFIDF2 で選択後の素性を確認すると，その多くはオランダ語から翻訳されていない語であった．DFD ドメインのテストデータも同じく，オランダ語から翻訳されなかった語を多く含んでいる．したがって，オランダ語の単語が素性として残ってしまったために，DFD ドメインでは性能が向上し，IFN ドメインでは悪化したと考えられる．

参考までに，翻訳ありと翻訳なしで，DFD のインドメインでのラベル付与性能を表 5 に示す．F1 に大きな変動はないが，Precision と Recall で大きな差が見られた．翻訳なしでは Precision が高く，Recall が低かったが，翻訳ありは逆になった．さらに Hamming Loss が大きく増加した．機械翻訳にて，尤度が最大化されるように英語に翻訳された結果，一般的な語が増加し，そして，より多くの素性が発火するようになったために，Recall が上昇したと考えられる．

6 おわりに

本稿では，異言語の資源を用いた物語の自動ラベル付与を提案した．モチーフインデックスを自動付与する問題をマルチラベル問題と考え，Okapi-BM25 と One-Versus-Rest 分類器を用いたモデルを作成した．Okapi-BM25 を用いたモデルでは Average Precision が 44.5%，One-Versus-Rest 分類器を用いたモデルでは F が 13%の精度を記録した．Okapi-BM25 によるランキングでは，WordNet を用いて同義語と上位概念語を追加した場合に，有意な精度向上がみられた．One-Versus-Rest 分類器による分類では，*bag-of-words* 素性では DFD のドメインに特有な分類であった．しかし，TFIDF 値による素性選択を行い，特徴語のみを素性とした場合は，IFN に汎化し，精度の向上が見られた．

オランダ語から英語への翻訳では，オランダ語の単語が多く残っていた．その結果，IFN への汎化が難しかったと思われる．そこで，今後は翻訳されなかったオランダ語単語への対処を行う．

表 4: One-Versus-Rest 分類器の実験結果（BOW が *bag-of-words* 素性，H. Loss が Hamming Loss）

テストコーパス	素性	正規化項	H. Loss	Subset Acc.	Acc.	Prec.	Rec.	F1
IFN	BOW	L1	0.06	0.06	0.09	0.09	0.09	0.09
		L2	0.06	0	0.03	0.06	0.04	0.03
	TFIDF1	L1	0	0	0	0	0	0
		L2	0.19	0	0.1	0.1	0.26	0.13
	TFIDF2	L1	0.07	0.06	0.06	0.06	0.06	0.06
		L2	0.07	0	0.03	0.06	0.03	0.04
DFD	BOW	L1	0.08	0.06	0.27	0.34	0.5	0.35
		L2	0.07	0.2	0.33	0.41	0.45	0.38
	TFIDF1	L1	0	0	0	0	0	0
		L2	0.14	0.2	0.27	0.3	0.4	0.31
	TFIDF2	L1	0.08	0.2	0.41	0.52	0.58	0.49
		L2	0.06	0.26	0.41	0.48	0.52	0.46

表 5: DFD における翻訳前と翻訳後の分類結果（*bag-of-words* 素性のみ，正規化項は L2）

テストコーパス	H. Loss	Subset Acc.	Acc.	Prec.	Rec.	F1
オランダ語	0.03	0	0.34	0.66	0.34	0.44
英語翻訳後	0.43	0.33	0.44	0.5	0.47	0.47

また，単語の出現に対する素性でなく，単語間の関係の素性を使用する予定である．物語分類においては「だれ」が「何をした」という情報が分類の判断に大きく寄与しており，したがって構文解析による統語的な素性を用いることで，単語の出現だけでは捉えられなかった情報を明確に示せるのではないかと期待される．

謝辞

本研究を行うにあたり，データの提供をしていただき，また考察のアドバイスを下さった，大阪大学・竹原新准教授にこの場をかりて感謝の意を表します．

参考文献

[1] Rudolf M. Dekker and Lotte C. Van De Pol. *The tradition of Female Transvestism in Early Modern Europe*. Palgrave Mecmillan, 1997.

[2] Folgert Karsdorp and Antal van den Bosch. Identifying motifs in folktales using topic models. pp. 41–49. In Proceedings of the *22 Annual Belgian-Dutch Conference on Machine Learning*, 2013.

[3] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[4] Graham Neubig, Kevin Duh, Masaya Oguchi, Takamoto Kano, Tetsuo Kiso, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. The naist machine translation system for iwslt2012. pp. 54–60. In Proceedings of *International Workshop on Spoken Language Translation*, 2012.

[5] M.T. Pilevar, H. Faili, and A.H. Pilevar. Tep: Tehran english-persian parallel corpus. pp. 68–79. In Proceedings of *12th International Conference on Intelligent Text Processing and Computational Linguistics*, 2011.

[6] Robertson S. and Zaragoza H. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, Vol. 3, pp. 333–389, 2009.

[7] Meder T. From a dutch folktale database towards an international folktale database. *Fabula*, Vol. 51, pp. 6–22, 2010.

[8] Stith Thompson. *Motif-Index of Folk-Literature: A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux*. Indiana University Press, 1955-58.

[9] 竹原新．ペルシア語口承文芸資料のデータ整形．イラン研究, Vol. 1, pp. 87–98, 2005.

[10] 樋口淳．民話の森の歩き方．春風社, 2011.

[11] 柳田国男．日本の昔話．新潮社, 1983.