Topic Modeling

トピックモデリングを活用した

北朝鮮の新年演説分析

KIM SUYEON

Contents

目次

Background

分析背景

北朝鮮の新年演説は、その年の政策方針や主要課題を示す重要な演説であり、
経済、外交、軍事などの分野において政府の意図を読み取る手がかりとなる。
しかし、以下の2つの問題がある。

1）演説は長文で専門的な表現が多いため、
　　政策の変化を体系的に分析することは容易ではない。

2）北朝鮮特有の政治用語や表記の違いにより、
　　従来の韓国語形態素解析器（KoNLPy など）が適切に機能しない。

Goal

分析目標

- ✓ 北朝鮮に特化した用語辞書を構築し、soynlpやMeCabなどの異なる形態素解析手法を組み合わせることで、トピック・モデリングの精度を向上させる。

- ✓ 直近12年間の北朝鮮の新年演説に繰り返し登場する主要なテーマを探索する。

  → 金正恩政権下での新年演説におけるテーマの変遷やトレンドを把握し、時期ごとに強調された内容の特徴を分析することを目的とする。

Dataset

データ説明

- 2012年1月1日から2024年1月1日までに発表された北朝鮮の新年演説の全文をデータセットとして使用する。
  → 合計13の文書で構成された、各年の新年演説のテキストデータであり、
  韓国語で記述されている。

- 各文書のテキストの長さは年度によって異なり、政治・経済・社社会に関する内容が含まれている。

Preprocessing

前処理

直面した問題

Problem1.スペース（分かち書き）の問題

KoNLPyライブラリが提供するHannanum,、Kkma、Komoranなどの形態素解析を利用する方法がある。これらはスペース（分かち書き）補正性能 も備わっているが、そもそも形態素を正確にトークン化できないため、この機能を活用することは難しい。分かち書きを補正するには、別の前処理手法を導入する必要がある。

Problem2.韓国語との違い

例えば、韓国語では「노선(ノソン）」という単語が、北朝鮮語では「로선（ロソン）」と表記される。そのため、形態素解析の際に「로（ロ）」と「선（ソン）」に分かれてしまい、単語を正しくトークン化できない問題が発生する。また、政治機関に関する用語の使い方も大きく異なり、同じ機関を指す表現が複数存在する場合が多い。この機能を活用することは難しい。分かち書きを補正するには、別の前処理手法を導入する必要がある。

Problem3.名前の抽出問題

特定の人物（例：指導者、歴史上の人物）や機関名が頻繁に登場する。これらの名前は新年演説のメッセージがどの点に集中しているかを把握する重要な手かがりとなるため、正確に抽出する必要がある。

Preprocessing

前処理

辞書の定義

4種類の辞書を定義する

Solution1. 権力構造と政治形態に関する辞書

北朝鮮における権力構造や政治形態を表す用語には、韓国とは異なる点が多い。北朝鮮政治ポータルが提供する権力構造と政治形態に関する説明を参考にし、辞書を作成した。

Solution2. ストップワード辞書および同義辞書

1-1）以下のサイトに掲載されているストップワード辞書を、現在のデータセットに合わせて修正した。
1-2）1文字の単語は誤分類される可能性があるため、2文字以上の単語のみを定義した。
2-1）1つの機関を指す用語が複数あることが確認された。
2-2）北朝鮮の機関について詳しくは分からないため、可能な限り把握している範囲で同義語辞書を定義した。

＊参考サイト： https://gist.github.com/spikeekips/40eea22ef4a89f629abd87eed535ac6a#file-stopwords-ko-txt

Solution3. 北朝鮮用語辞書および同義語辞書の活用

1）1文字の単語は誤分類率を高めるため、2文字以上の単語のみを使用する。
2）動詞も含まれているため、「～다」で終わる単語は削除する。
3）文も含まれているため、その部分は削除する。

＊参考サイト：https://nkinfo.unikorea.go.kr/nkp/word/nkword.do

Preprocessing

前処理

トークン化

単語のトークン化をどのように進めたのが、説明する。
＊各段階で抽出された名刺はデータセットから削除し、次の階段へ進む。

Step1. 特殊文字の削除

Step2. 名前の抽出後、同義語の処理

Step3. 権力構造・政治形態に関する辞書を使用し、名詞を抽出後、同義語の処理

Step4. ストップワード（Stopwords）の除去

Step5. soynlp名詞抽出器を使用し、名詞を抽出後、同義語の処理

Step6. 北朝鮮用語辞書を活用し、名詞を抽出後、同義語の処理

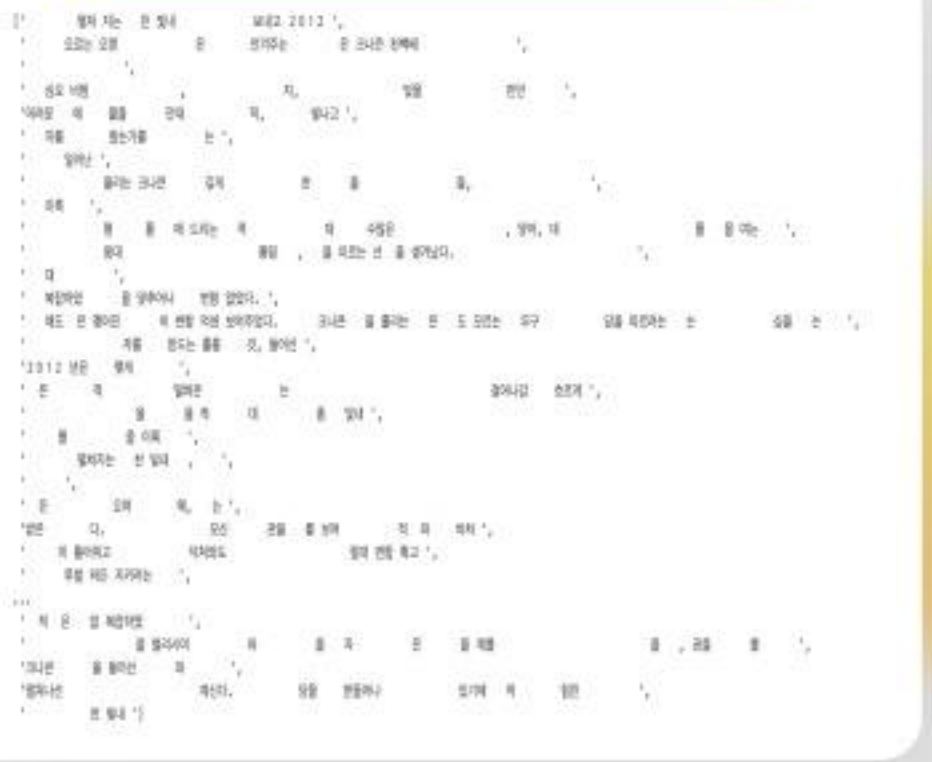
Step7. MeCab名詞抽出器を使用し、名詞を抽出後、同義語の処理

Preprocessing

前処理の結果

トークン化前

トークン化後



Modeling

モデリング

- **LDAモデル（gensimライブラリの活用）**
  - Coherenceの観点から、TF-IDFやn-gramを使用しないモデルのCoherence値の方が高かったため、抽出した単語のみを使用してLDA（Latent Dirichlet Allocation）ベースのトピックモデリングを実施した。
  - gensim.models.ldamodel.LdaModelを使用し、トピック数や学習パラメータを調整しながら最適なLDAモデルを構築した。
  - 前処理された文書-単語行列を基に学習を実施した。

- **CoherenceModelモデルを用いたモデル評価**
  - gensim.models.CoherenceModelを活用し、トピック間の一貫性（CV coherence）スコアを計算。最も一貫性の高いトピック数を選定し、最終モデルを決定した。

- **pyLDAvisによるトピックの可視化**
  - pyLDAvis.gensim_modelsを使用し、各トピックの主要単語やトピック間の関係を可視化した。

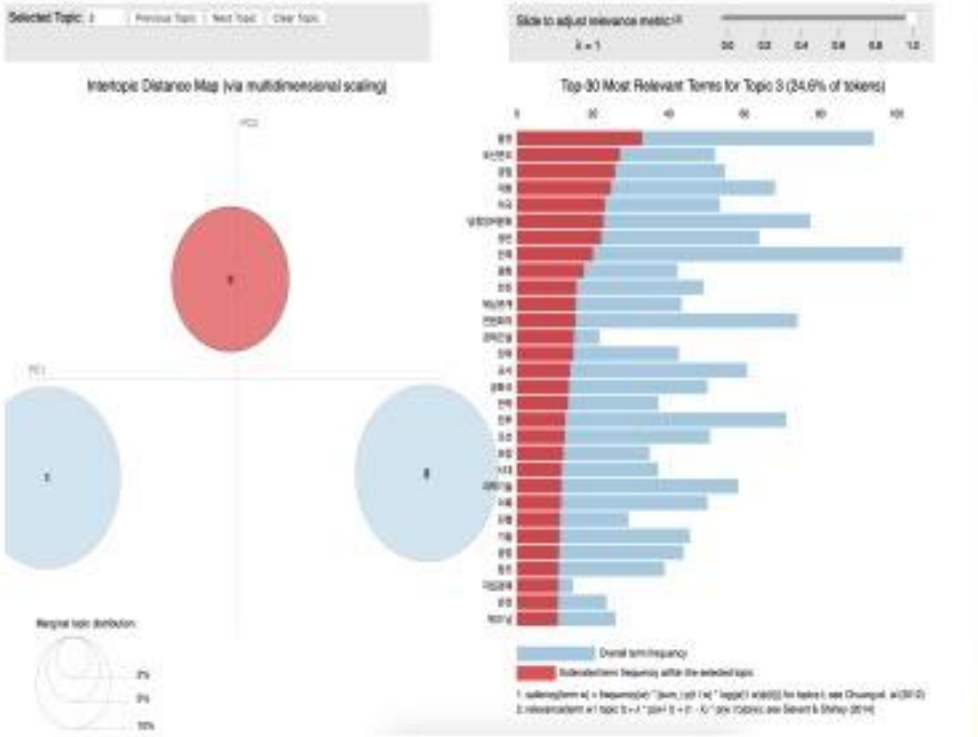
Modeling

最適なトピック数：３

- Coherenceスコアは0.6322413885329 32で最も高かった。
- ３つのトピックすべてにおいて、推定された単語の頻度が異なっていた。Lambda値によって単語の順位が変動する。

トピックを構成する主要単語

- トピック１（国防）：전투（戦闘）、군대（軍隊）、북남관계（南北関係）、평화（平和）、전쟁（戦争）
- トピック２（経済・社会）：농촌（農村）、농업（農業）、과학기술（科学技術）、문화（文化）
- トピック３（政治）：당중앙위원회（党中央委員会）、미국（アメリカ）、군사（軍事）、과학기술（科学技術）、자립경제（自立経済）



Result

分析結果

2012年~2017年, 2018年~2021年, 2022年~2024年の3つの期間に分けて、トピックの変遷について説明する。

- **2012年~2017年**
  - 政治が主要なトピックであった。
  - 2012年、金正恩が政権を握り始めた年は、「政治」の割合が非常に大きかった。
  - 2012~2017年のうち、特に政権初期の2012年に「政治」の割合が突出して高かった点は、当時の政権安定化や体制構築のための政策方針と意志を強く反映していると解釈できる。

	国防	経済および社会	政治
2024	0.001637	0.997062504	0.001301
2023	0.004122	0.992393434	0.003485
2022	0.001401	0.997196376	0.001403
2021	0.72594	0.256100237	0.01796
2020	0.762569	0.235289365	0.002142
2019	0.991454	0.002423041	0.006123
2018	0.970676	0.002134206	0.02719
2017	0.181237	0.003784428	0.814979
2016	0.005335	0.002920977	0.991744
2015	0.003318	0.002403126	0.994279
2014	0.003725	0.003946392	0.992329
2013	0.00243	0.002192244	0.995378
2012	0.001941	0.002019707	0.996039

Result

分析結果

- **2017年~2021年**
  - 2017年末から深刻化した対北朝鮮制裁の影響により、この期間では「国防」が主要なトピックとなった。
  - 2020年からはコロナ禍で「経済および社会」に関するトピックの割合が増加し始めた。

- **2022年~2024年**
  - コロナ以降、「経済および社会」が主要なトピックとなった。
  - コロナ禍による経済難が深刻化したことで、2022年の「経済および社会」トピックの割合が最も高くなった。

	国防	経済および社会	政治
2024	0.001637	0.997062504	0.001301
2023	0.004122	0.992393434	0.003485
2022	0.001401	0.997196376	0.001403
2021	0.72594	0.256100237	0.01796
2020	0.762569	0.235289365	0.002142
2019	0.991454	0.002423041	0.006123
2018	0.970676	0.002134206	0.02719
2017	0.181237	0.003784428	0.814979
2016	0.005335	0.002920977	0.991744
2015	0.003318	0.002403126	0.994279
2014	0.003725	0.003946392	0.992329
2013	0.00243	0.002192244	0.995378
2012	0.001941	0.002019707	0.996039

Conclusion

プロジェクトの意義

- 新年演説を分析することで、北朝鮮の政治的・経済的方向性を客観的に把握できる基礎資料を提供し、国際関係構築に必要なインサイトを示した。つまり、北朝鮮の政策的方向性を予測することに貢献できた。

- 北朝鮮で使っている言語みたいな特殊なテキストデータにおいて、トピックモデリング手法の活用可能性を検証した。これにより、他の言語的特殊性を持つデータにもトピックモデリング手法を応用できることを示した。