表層情報と深層情報による半教師あり学習を用いた重要文抽出システム

天野　禎章　　横山　晶一

山形大学大学院理工学研究科

1．研究概要

我々は、人手作成により近い要約を生成するシステムに取り組み、特に単一テキスト要約で高精度なシステムの構築に力を入れている[1]。近年の自動要約研究の多くは複数テキストを対象としており、英文をメインに扱った評価型ワークショップ DUC(Document Understanding Conference[2], 現在は TREC(Text Retrieval Conference)[3]の QA トラックと統合して TAC(Text Analysis Conference)[4]のトラックの一つへ移行)や日本語を入力言語とした TSC(Text Summarization Challenge)[5]でも複数文書を扱っている。しかしながら複数テキスト要約では、要約率や冗長性などを考慮する必要性があり[6]、各入力文書集合から重要文抽出する場合にもより高い精度が要求される。よって、高精度な複数テキスト要約システムを構築するには、少なくとも単一テキスト要約システムで充分な成果を得る必要がある。

自動要約システムは、単語頻度や出現位置などの表層的な情報を利用するアプローチと、辞書やシソーラスなどの知識や特殊な解析を要する深層的な情報を利用するアプローチがあり、前者は頑健性を持つ一方で精度の向上が難しく、後者は高い精度が期待できる一方で汎用性が乏しい。また、両方の情報を利用したシステムでは精度は向上するが、汎用性が低下する。重要文抽出は文の重要・非重要の二値分類に言い換えられ、分類タスクでは教師あり学習による手法が優れた成果を得ている。教師あり分類の精度は、一般に有効な素性と教師ラベル付きデータの量に依存し、人手による教師ラベルデータを増やすことで向上が期待できる。しかしながら、教師ラベルの付与は時間的コストがかかる。そのため、ラベルありデータとラベルなしデータを混在させた半教師あり学習で分類器を構築する手法が提案され、自動要約に適用した研究[7,8]もある。

以上のことから本研究では、頑健性を保ったまま精度を向上させる方法として、表層情報と深層情報からそれぞれ独立した二つのシステムを構築し、半教師あり学習を行った。具体的には、表層情報とそれらを基にしたグラフベースのスコアを素性とした教師あり学習による要約器と、日本語 WordNet[9]や日本語語彙大系[10]などの知識情報と主題・焦点[11]や LDA(Latent Dirichlet Allocation)[12]などの解析を介した結果を基にしたグラフベースのスコアを素性とした教師あり学習による要約器を用いて、Co-training[13]を行った (先行研究[8]では、同一の素性ベクトルから異なる学習アルゴリズムによる分類器を用いて、Co-training を実行している)。その結果、要約率 50%のとき両システムともに F 値が向上した。

2．素性抽出

自動要約の分野で利用される情報は、単語頻度や文書での位置などの表層情報を始め非常に多く提案され、有効性が論じられている[14,15]。本研究では、先行研究を参考に有効な素性を抽出し、機械学習ベースの要約システムを構築した。

システムでは、まず入力文書を TSC のフォーマットに変換し、各文を Mecab(Version 0.97)[16]を用いて形態素解析する。次に各文を分かち書きした単語集合から表層情報と深層情報を抽出する。そして、これらの各情報からそれぞれグラフを作成し、ネットワーク分析で用いられる指標やランキングアルゴリズムによるスコアをグラフベースの素性として利用する (下図)。このとき、深層情報ベースの素性は一部を除きグラフベースの素性へと「変換」するが、表層情報ベースの素性では一部を除きグラフベースの素性を「追加」する。

これらのステップを学習データ集合に対して行い、最終的には各文の素性ベクトルを用いて分類器の学習あるいは文の分類を行う。

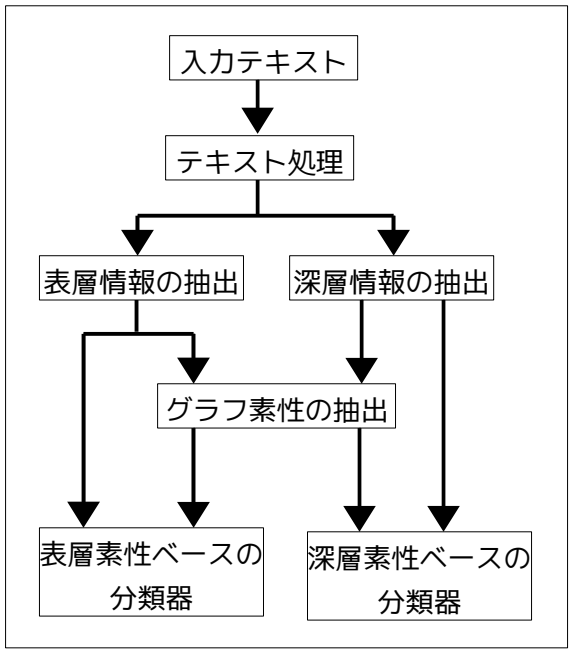

図　入力から分類器学習・文の分類までの流れ

2.1 表層情報による素性

表層情報ベースの素性は、単語頻度と位置情報などに基づき次の九個とした。

1) 正規化文字数：文中の文字数÷全文字数
2) 頻度率：文中の頻度の和÷単語総数
3) 最小出現位置率：正規化最小出現位置÷文中の内容語数
4) 単語率：文中の単語数÷単語総数
5) 内容語率：文中の内容語数÷文中の単語数
6) 異なり単語数：文中の異なり単語数÷全異なり単語数
7) 異なり単語率：文中の異なり単語数÷文中の内容語数
8) 各文の単語頻度：文中の単語出現頻度
9) 文全体での単語頻度：文中の単語の文全体での出現頻度

これらを抽出後、8)と 9)のそれぞれと、1)から 7)の素性からグラフを作成し、2.3 節のグラフベースの素性を求める。最終的な表層情報ベースによる要約システムは、次の素性を統合したベクトルを用いた分類器となる。

- 1)から 7)の表層情報による素性
- 1)から 7)の素性によるグラフベース素性
- 8)の素性によるグラフベース素性
- 9)の素性によるグラフベース素性

2.2 深層情報による素性

深層情報ベースの素性は、辞書やシソーラスなどの知識情報と解析システムを介する情報に基づき次の八種類計十三個とした。

1) 日本語 WordNet:上位語 synset の頻度。
   下位語 synset の頻度。
   上記を含む全意味的関連の synset の頻度。
2) 日本語語彙大系：意味属性番号の頻度。
   語彙大系木の根までの部分経路の頻度(類似度算出に利用(Subpath Set[17]))。
3) 単語感情極性対応表[18]：文中の内容語のうち、正極範囲内にある単語のスコアの和。
   文中の内容語のうち、文中の内容語のうち、負極範囲内にある単語のスコアの和。
   正極値と負極値の和。
4) 係り受け解析結果：CaboCha[19](Version 0.60 pre4)による係り受け解析木の根までの部分経路の頻度(Subpath Set で類似度)。
5) 固有表現：八種類タグ (CaboCha で抽出) の文書中での頻度平均。
6) 主題・焦点：全内容語に対する主題と焦点 (Conditional Random Field(CRF)[20]を用いて抽出) の頻度 (内容語数次元のベクトルで、判定された語を加算) 。
7) 文タイプ[21]：符号に基づき二値分類器を組み合わせたマルチラベル分類器によって自動付与した各ラベルの頻度。
8) LDA トピック：文中に含まれる内容語の各トピック確率の和。モデル作成には『現代日本語書き言葉均衡コーパス』モニター公開データ (2009 年度版) [22]を利用し、トピック数は 50 とした。

これらを抽出後、3)と 5)を除外した各素性からそれぞれグラフを作成し、次節で説明するグラフベースの素性に変換する。最終的な深層情報ベースによるやくシステムは、次の素性を統合したベクトルを用いた分類器になる。

- 3)と 5)の深層情報による素性
- 3)と 5)を除く素性によるグラフベース素性

2.3 グラフベースの素性

グラフベースの素性抽出は、1)それぞれの素性情報から複数の観点に基づくグラフを作成し、2)ネットワーク分析で用いられる指標やランキングアルゴリズムなどから各ノード (文) のスコアを算出する、二つの工程を経る。

第一段階として、抽出した表層情報と深層情報の各素性からコサイン類似度グラフ (重み付き無向グラフ) を作成する (表層情報と深層情報) 。このとき、非類似性を考慮するため、類似度が閾値内 (0 を除く 0.5 以下) のノード間をリンクさせる無向グラフ (表層情報と深層情報) と、隣接よりも遠くを考慮するため、閾値設定 (1 を除く 0.5 以上) した類似度グラフの共通隣接ノード数を要素とした行列に拡散カーネル[23]を適用した重み付き有向グラフ (表層情報は隣接で充分として、深層情報のみ) を作成した。

このように作成された三つないしは二つのグラフから次数中心性や近接中心性、パワー中心性 (パラメータは 0.5) と拘束性、そしてランキングアルゴリズムの SALSA(The Stochastic Approach for Link-Structure Analysis)[24]による HUB 値と AUTHRITY 値の和を求め、またタイトルとの類似度、最初の文との類似度、各パラグラフの最初の文との類似度、類似度グラフの主固有ベクトルを素性として加えた。なお、一部のグラフで除外したグラフベース素性もある (例えば、リンクが著しく少ない傾向にあるグラフでの近接中心性) 。

3. 実験

前述した素性を用いて分類器を学習する。予測ラベルを付与する分類器の学習アルゴリズムは、Random Forest[25]を利用した。Random Forest は、多数の決定木を用いた集団学習アルゴリズムで、与えられたデータセットからブートストラップサンプルを作成し、各サンプルデータを用いて未剪定の決

定・回帰木を生成 (ランダムに選択した素性の中で CART の Gini 係数や ID3 の情報量利得などの規準によって最善のものを利用してノードを分岐)、全ての結果を統合し新しい分類器を構築する。非常に多くの素性を扱え、さらに学習速度が早く、また精度が高いという特長がある。

ラベルなしデータは、NTCIR[26]で提供されているテストコレクションに含まれる記事と要約生成に適さないテキスト (新聞休止や人事の知らせ、文書サイズが閾値よりも小さい) を除いた 94,95,98 年の毎日新聞データ集合からを合わせて 3000 文書をランダムに抽出し、分類器に基づき予測ラベルをつけた。このとき、ラベルの確信度の高い文を要約率を満たすまで貪欲法で選択し、選ばれた文を正例とした。この予測ラベル付きデータを用いて再度分類器を学習し、同じように抽出した 3000 文書に対して予測ラベルをつけるというステップを繰り返し、のべ 9000 文書に予測ラベルを付与した。

4. 評価

教師ありデータに予測ラベル付きデータを足したとき、システムの精度がどれだけ変化したか検証するため、10 分割交差検定を行った。このとき、教師ありデータのみを分割し、予測ラベル付きデータは学習データに全て追加して評価した。

そのときの各モデルを用いた要約で F 値を算出した結果が次の表である。CR は指定要約率を表し、ベースラインには、表層情報ベースのシステム (SystemA)と深層情報ベースのシステム(SystemB)、表層情報と深層情報の両方の素性を利用した要約システム(SystemC)の 10 分割交差検定による結果を提示した。また、Co-training による予測ラベルを追加して学習した結果が Co-SystemA と Co-SystemB であり、括弧内は教師あり学習時との差である。

高精度を期待した深層情報ベースの結果が低く、要約率 50%になって鍛えたい表層情報ベースを越えた。この点から要約率が低い要約では知識情報よりも頻度や位置を、要約率が高い要約では深層情報を、重視する傾向があると思われる。だとすると、要約率自体も有効な素性となりうるだろう。あるいは、入力テキストサイズが小さい場合、必然的に要約文も少なく、人が選ぶ文が頻度や位置情報に依る傾向があるかもしれない。

表　10 分割交差検定したときの F 値の平均

	CR10%	CR30%	CR50%
SystemA	0.443	0.492	0.618
SystemB	0.320	0.453	0.631
SystemC	0.397	0.477	0.633
co-SystemA	0.376 (-0.067)	0.480 (-0.012)	0.647 (**+0.029**)
co-SystemB	0.331 (**+0.011**)	0.470 (**+0.017**)	0.644 (**+0.013**)

直観的に SystemA の精度を上げるには、少なくとも SystemB がそれ以上の精度である必要があるが、その前提を守れた要約率 50%では、SystemB の精度は落ちずに両システムともに精度が向上している。この現象は要約率 10%と要約率 30%では生じていない。この点について全素性を利用した SystemC に着目すると、要約率 10%と要約率 30%の場合は SystemA と比べて精度は下がっているが、要約率 50%では上がっている。素性を分割せずに教師あり学習したシステムが分割して教師あり学習した二つのシステムよりも高い精度のとき (分割した素性でそれぞれ学習した分類器において、抽出する正解文が重複する割合が低い場合と思われる)、両システムの精度が向上する可能性がある。この予想に基づくと、要約率が低い要約では両システムの向上は難しく、少なくとも片方が優れている必要がある。

5. まとめと今後の方針

本稿では、頑健性を持たせたままシステムの精度を向上させる方法として、高精度が期待できる深い解析や知識ベースによるシステムとの Co-training でラベルなしデータに予測ラベルをつけた半教師あり学習による要約システムを構築した。その結果、両システムの精度が向上した。今回の実験では、一度にラベル付けするデータ数を 3000 文書とし、繰り返し回数を２回としたが、文書の総数を増やすことで、さらなる向上が期待できる。しかし、学習データを一括に与えるバッチ学習ではデータ数が増えることで消費メモリと学習に要する時間が膨大になる。よって、学習データを増やす場合は逐次的に処理するオンライン学習の利用が望ましい。

深層情報ベースのシステムの精度向上は、表層情報ベースに比べて制限が緩く大きな向上が望める。言論マップコーパス[27]を用いたテキスト含意関係の考慮、格情報や助詞の利用、主題・焦点の抽出システムと文タイプのラベル付けのシステムの精度の向上、不要な素性の除去などの深層情報自体の処理工程以外に、グラフベースの素性に変換時の最適なパラメータ推定や閾値設定などを行うことで低い要約率でも Co-training による表層ベースのシステムの精度向上に役立てられる。

また、表層情報ベースで起点となる数文を抽出した後、深層ベースから求めた文間の類似度やリンク確率で MMR(Maximal Marginal Relevance)[28]のような動的スコアを算出する方法も考えられる。

精度が充分に上がったシステムを複数テキスト要約や単語・文節を単位にした文短縮への適用、言語資源が豊富な言語との言語横断要約システムなどの実験を行っていく。

謝辞

NTCIR より提供されるテストコレクションがシステムの構築と評価において大きく貢献しました。NTCIR に多大な感謝を申し上げます。

参考文献

[1]天野禎章，横山晶一，“不要文除去を目的とした重要文抽出システム”，言語処理学会第 15 回年次発表論集，pp.64-67，2009.
[2]TREC，“http://trec.nist.gov/”(アクセス:2010.1.7)
[3]DUC，“http://duc.nist.gov/”(アクセス:2010.1.7)
[4]TAC，“http://www.nist.gov/tac/”(アクセス:2010.1.7)
[5]TSC，“http://lr-www.pi.titech.ac.jp/tsc/”(アクセス:2010.1.7)
[6]奥村学，難波英嗣，“テキスト自動要約”，オーム社(2005)
[7]Massih-Reza Amini, Patrick Gallinari, “The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization”, Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.105-112, 2002.
[8]Kam-Fai Wong, Mingli Wu, Wenjie Li, “Extractive Summarization Using Supervised and Semi-supervised Learning”, Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1, pp.985-992, 2008.
[9]Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, Kyoko Kanzaki, “Enhancing the Japanese WordNet”, Proceedings of the 7th Workshop on Asian Language Resources, pp.1-8, 2009.
[10]池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦，“日本語語彙大系”，岩波書店(1997)
[11]吉田悦子，横山晶一，“主題・焦点を用いた文脈解析の一手法”，電子情報通信学会技術研究報告 97(330)，pp.1-8，1997.
[12]D.Blei, A.Ng, M.Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, pp.993-1022, 2003.
[13]Avrim Blum,Tom Mitchell, “Combining Labeled and Unlabeled Data with Co-Training”, Proceedings of the 11th Annual Conference on Computational Learning Theory, pp 92-100, 1998.
[14]H.P.Edmundson, “New methods in automatic abstracting”, Journal of the Association for Computing Machinery 16(2), pp.264-285, 1969.
[15]Ani Nenkova, Lucy Vanderwende,Kathleen McKeown, “A Compositional Context Sensitive Multidocument Summarizer: Exploring the Factors That Influence Summarization”, Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.573-580, 2006.
[16]奈良先端科学技術大学院大学，MeCab version 0.98.
[17]箱田慶太，市川宙，本泰一，徳永健伸，“構文的類似度を用いた文の検索”，言語処理学会第 12 回年次発表論文集，pp.1131-1134，2006.
[18]高村大也，乾孝司，奥村学，“スピンモデルによる単語の感情極性抽出”，情報処理学会論文誌ジャーナル，Vol.47 No.02 pp.627-637，2006.
[19]工藤拓，松本裕治，“チャンキングの段階適用による日本語係り受け解析”， 情報処理学会論文誌，Vol.43,No.6，pp.1834-1842，2002.
[20]John D.Lafferty, Andrew McCallum, Fernando C.N.Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, Proceedings of the Eighteenth International Conference on Machine Learning, pp.282-289,2001.
[21]関洋平，“ジャンルとテキスト構造に着目した自動要約”，総合研究大学院大学 博士論文，2005.
[22]国立国語研究所，『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）
[23]Risi Imre Kondor, John D. Lafferty, “Diffusion Kernels on Graphs and Other Discrete Input Spaces”, Proceedings of the Nineteenth International Conference on Machine Learning, pp.315-322,2002.
[24]R.Lempel, S.Moran, ”SALSA: the stochastic approach for link-structure analysis”, ACM Transactions on Information Systems, Vol.19,No. 2, pp.131-160, 2001.
[25]Leo Breiman, “Random Forests”, Machine Learning, vol.45, pp.5-32, 2001.
[26]NTCIR,“http://research.nii.ac.jp/ntcir/index-ja.html”（アクセス:2010.1.7）
[27]村上浩司，松吉俊，乾健太郎，松本裕治，“言明間の意味的関係の体系化とコーパス構築”，言語処理学会第 15 回年次発表論文集，pp.602-605，2009.
[28]J.Carbonel, J.Goldstein, “The use of MMR, diversity-based reranking for reordering documents and producing summaries” Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.335-336,1998.