ラベル伝播による他言語資源の利用と転移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
{tdy56650,yokoyama}@{st,yz}.yamagata-u.ac.jp

1　研究概要

多くの自動要約に用いられる情報は、次の三つに大別できる。

- 表層情報ベース (Surface-based features):
  単語頻度、位置情報など
- 知識情報ベース (Knowledge-based features):
  辞書やシソーラス、手がかり語など
- 解析情報ベース (Analysis-based features):
  照応解決や文書構造解析、含意関係認識など

これらの情報は、言語に依存しない表層情報[1](表層情報ベース) と、言語に依存する深層情報[2](知識情報ベースと解析情報ベース) に分けられる。前者は、他言語に適用可能で頑健性がある一方で、利用できる情報が乏しく一定以上の精度向上が期待できない。対して後者は、意味的関連性や一貫性などを考慮可能で高精度が期待できる反面、言語に大きく依存する。

本研究では、DBFを用いて頑健性を保ったまま精度向上する手法として、他言語のラベルなしコーパスに同じく他言語リソース (辞書やシソーラス) に基づく素性から関連度を求めてラベル伝播 [1, 2] させ、その結果を転移学習によってターゲット言語に適用した。転移学習 [3] は、「ある問題を効果的かつ効率的に解くために、別の関連した問題のデータや学習結果を再利用する」ことで、大量のラベルなしテキストの有効利用に半教師あり学習や能動学習と並んで研究されている機械学習の一枠組みである。集団学習に基づく手法を転移学習に拡張した研究 [4, 5, 6, 7] や、オンライン学習に基づく方法を導入した研究 [8] の他に、グラフとNormalized Cutを用いた研究 [9, 10] などがあり、多くの優れた成果が報告されている。提案手法は、ソースとターゲットともにラベルが存在する状況に相当し、言語リソースが乏しいターゲット言語でも機械翻訳を介さずにソース言語の豊かなリソースを転用できる。これにより、元々の結果よりもF値が大きく向上した。なお、言語リソースを使うソース言語は日本語、使わないターゲット言語は英語とした。

[1]SBF と表記。
[2]Detail-based features; DBF と表記。

2　システム概要

提案手法を用いた重要文抽出システムは、入力テキストを文レベルの素性ベクトルに変換する素性抽出モジュールと、素性情報からモデルの学習・分類を行う分類器モジュールの二つから構成されている。

2.1　素性抽出モジュール

素性抽出モジュールでは、各文単位でBOW形式に変換し (日本語テキストはTSC[3]フォーマットに基づく文単位に、形態素解析[4]と内容語抽出を実行する。英語テキストはDUC[5]のタグ付けPerlスクリプトで出力されたタグを一文単位として、Snowballステミングとストップワードを除去する)、SBFとDBFを要素とする素性ベクトルへ置き換える。この工程で算出する素性は、先行研究 [11] をベースに、CentroidとLexRank[12] や、ALAGINのデータベース[6]を用いたグラフ、Mcutによるグラフのグラフベース素性などを新たに追加した。

[3]Text Summarization Challenge. http://lr-www.pi.titech.ac.jp/tsc/ (June, 2010).
[4]Mecab Ver.0.98 pre3, http://sourceforge.net/projects/mecab/ (June, 2010).
[5]Document Understanding Conference. http://duc.nist.gov/ (June, 2010).
[6]http://www.alagin.jp (June, 2010).

2.2 分類器モジュール

分類器モジュールでは、前モジュールを経た素性データから、モデル作成済みの場合はモデルを用いて分類 (SBF のみ使用) を、モデル未作成時は学習を行う。モデルの学習には、ソース言語のラベルあり (SBF+DBF) とラベルなし (SBF+DBF) のデータ、ターゲット言語のラベルあり (SBF) の素性データが必要になる。

モデル学習の工程は、二段階の処理から成り立つ。始めに、ソース・ターゲットのラベルありデータを用いて、ソース言語のラベルなしデータにラベル伝播を行う。このとき、ソース言語では SBF と DBF を、ターゲット言語では SBF のみを用いる。次にラベルありデータと擬似ラベルが伝播されたデータの両方を同等に扱い、TrAdaBoost[4] の手法を適用する。共通語彙集合を介して元ドメインから目標ドメインに直接ラベルを伝播させる研究 [13] があるが、異言語間ではパラレルコーパスなしに共通語彙集合を作れず応用ができない。本手法では、言語非依存の SBF を介してソース言語のコーパスにラベルを伝播することで解決し、他言語コーパスの流用も可能になった。これに加えて、ソース言語の関連度の計算に DBF も利用することで事例間の類似度をより的確に図れる。そして、少なからず生じる SBF 間の言語差をドメインの違いとして捉え、転移学習にてこれを低減した。

モデル学習の工程は次の通りである。

必要データと設定 :

- ラベル (±1) ありデータ集合:
  ソース言語データ集合 LS = {(SBF,DBF)};
  ターゲット言語データ集合 LT = {(SBF)};
- ラベルなしデータ集合:
  ソース言語データ集合 US = {(SBF,DBF)};
- Learner = ベース学習アルゴリズム;
- N = 弱分類器数;

初期化と定義 :

重みベクトル $w^1 = (w_1^1, \ldots, w_{n+p+m}^1)$ を初期化,
$n = LS$ のデータ数,
$p = US$ のデータ数,
$m = LT$ のデータ数;

ループ for t = 1,..., N:

1. $P = \frac{w^t}{(\sum_{i=1}^{n+p+m} w_i^t)}$ を計算する;
2. P に基づきブートストラップサンプリングしてサンプル集合 X を次のように構成する;

$$X = \begin{cases} X_i^{LS}, & i = 1, \ldots, n_s; \\ X_i^{US}, & i = n_s+1, \ldots, n_{src}; \\ X_i^{LT}, & i = n_{src}+1, \ldots, n_{tgt}; \end{cases}$$

$X^{LS} = LS$ から抽出したデータ,
$X^{US} = US$ から抽出したデータ,
$X^{LT} = LT$ から抽出したデータ,
各々のデータ数を上から n_s, n_u, n_t とし,
$n_{src} = n_s + n_u, n_{tgt} = n_{src} + n_t$ とおく;

3. X^{LS}(SBF と DBF) と X^{US}(SBF と DBF) から関連度 Rd_i を, X^{LT}(SBF) と X^{US}(SBF のみ使用) から関連度 Rd_o を算出する;
   先行研究 [13] 同様に Rd_i と Rd_o を各行と列に対する次数対角行列を用いて正規化する;
4. X^{LS} と X^{US} のラベルを要素とする対角行列 Cd_i と, X^{LT} と X^{US} のラベルを要素とする対角行列 Cd_o を作成する (ラベルなしデータの初期要素は「0」を代入);
5. 式 (1) の固有ベクトル E_V を算出する;

$$(Rd_i)^T Cd_i Rd_i + (Rd_o)^T Cd_o Rd_o \qquad (1)$$

6. E_V の対角成分を符号関数 (Sign) に与えた戻り値を X^{US} の擬似ラベルとする;
7. 4 から 6 の処理を収束するまで繰り返す;
8. X^{US} の擬似ラベルが「0」のままのデータを X から除外する;
9. 学習データ集合 $T = \{(X, c(X))\}$ を用いて Learner によって弱分類器 h_t を構築する,
   $c(X)$ はサンプル集合 X に対応するラベルであり, [-1,+1] から [0,+1] に置き換える;
10. あるデータ X_i が与えられたときの h_t の分類結果を $h_t(X_i)$ としたとき,
    $Loss_i = |h_t(X_i) - c(X_i)|$ として X^{LT} に対するエラー率 ε_t を計算する (ただし, $\varepsilon_t \leq 0.5$ でなければならない);

$$\varepsilon_t = \sum_{i=n_{src}+1}^{n_{tgt}} \frac{w_i^t \cdot Loss_i}{\sum_{i=n_{src}+1}^{n_{tgt}} w_i^t}$$

11. 信頼度 β_t を下記の式で求める;

$$\beta_t = \frac{\varepsilon_t}{(1-\varepsilon_t)}, \quad \beta = \frac{1}{(1+\sqrt{\frac{2\ln(n)}{N}})}$$

12. 次式にて重みベクトル w_i^{t+1} を更新する;

$$w_i^{t+1} = \begin{cases} w_i^t \beta^{Loss_i}, & i = 1, \ldots, n_{src}; \\ w_i^t \beta_t^{-Loss_i}, & i = n_{src}+1, \ldots, n_{tgt}; \end{cases}$$

ターゲット言語の強分類器 (SBF にて推定) が生成:

$$H(x_i) = \begin{cases} 1, & \prod_{t=[N/2]}^{N} \beta_t^{-h_t(x_i)} \geq \prod_{t=[N/2]}^{N} \beta_t^{-\frac{1}{2}} \\ 0, & otherwise; \end{cases}$$

3 提案手法の実験

ソース・ターゲット間で他言語リソースを用いた本手法の有効性を判定するため、5 点交差検定によって F 値を求めた。このとき分類モデルの純粋な精度を図るため、要約率を満たすまで文を選択せずに分類結果 (出力 = +1, 非出力 = 0) で算出した。

ベース学習アルゴリズム (Learner) には Random Forest[14] を用いた。Random Forest は、与えられたデータセットからブートストラップサンプリングを作成し、各サンプルデータを用いて未剪定の決定・回帰木を生成、全ての結果を統合した分類器を構築する集団学習アルゴリズムである。多くの素性を扱え、SVM[15] と比べて学習速度が早く、また精度も高いという特長がある。設定する弱分類器数において、TrAdaBoost では学習誤り率の収束の観点から 50 個以上を推奨しているが、提案手法ではラベル伝播工程で大きく時間を消費するため 20 個とした。また、同じく学習時間を考慮してラベル伝播時に収束するまで繰り返さず、一度の伝播のみで実験した。重みベクトル w^1 は「1」にて初期化し、ラベル伝播時に作成するグラフの関連度 Rd_i と Rd_o は RBF カーネル ($\sigma = 1$) によるグラム行列 (x_i, x_j のカーネル値 $K(x_i, x_j)$ を i, j 成分とする行列) を用いた。

本手法は、ターゲット言語の分類器の精度向上をソース言語のラベルなしデータ数 n_u を増やすことで期待できる。しかしながら、ラベル伝播時に行列と固有値計算を伴うため、大規模な記録領域を要する。これを避けるため、ラベルなしデータ US を L 個に分割し ($US_1, US_2, \ldots, US_L$)、$L \times N$ 個の弱分類器を構築して強分類器 $H2 = \{H_1, H_2, \ldots, H_L\}$ を構築した (本稿では $L = 5$ とした)。また、強分類器 $H2$ から信頼度 β_t に準じて N 個の弱分類器を選別して構築した強分類器 $H3$ も併せて実験した。

3.1 データセット

実験に用いたデータセットは次の通りである。

- **テストデータ**:DUC2001 の単一テキストを対象にした 100 単語要約生成タスクを、重要文抽出用に変換したデータ。DUC のデータに提供されている Perl スクリプトを利用して作成した。
- **学習データ**:交差検定時の評価用データを除くテストデータ (ターゲット言語) と、NTCIR[7]による単一テキストを対象とした重要文抽出タスク (TSC-1 の A-1 の要約率 50%) のテストコレクション (ソース言語)。
- **ラベルなしデータ**:毎日新聞コーパス (94,95,98,99) からランダムに選択した 3000 記事 (約 50000 文)。ただし、サンプリング時に要約生成に適さない文書 (休刊日の通知、人事の知らせ、短すぎる記事 (1KB 未満)) とテストコレクションを除いている。

[7]NII-NACSIS Test Collection for IR Systems. http://research.nii.ac.jp/ntcir/index-ja.html (June, 2010)

表 1: F 値による評価結果

	F-measure	Precision	Recall
BSLineTgt	0.23179	0.53644	0.14892
BSLineSrcTgt	0.21770	0.43490	0.14642
LbPropDirect	0.32974	0.20228	0.89139
LbPropInDirect	0.28450	0.21793	0.41223
TrAdaBoost(20)	0.26063	0.41090	0.19084
TrAdaBoost(50)	0.26301	0.38128	0.20363
TrAdaBoost(100)	0.26384	0.39068	0.20313
TrAdaBoost(200)	0.27525	0.40543	0.21239
TrAdaBoost(500)	0.28224	0.40361	0.22151
OurMethodH2	**0.34767**	0.26994	0.50084
OurMethodH3	0.33889	0.27831	0.44491

3.2 実験結果

提案手法による重要文抽出システムを評価した結果が、表 1 である。ベースラインとして、テストデータ (SBF) で学習したモデル (BSLineTgt), テストデータにソース言語のラベルありデータ (SBF) を追加して学習したモデル (BSLineSrcTgt), ソース言語の学習データを用いてラベル伝播法でテストデータにラベルを伝播 (LbPropDirect), ソース言語のラベルなしデータにラベル伝播した結果を追加して学習したモデル (LbPropInDirect),TrAdaBoost の弱分類器数を変動させたモデルを提示した。このとき、LbPropDirect はソース・ターゲット間の SBF のみで関連度を計算し、先行研究 [1] の手法によってラベル伝播させている。LbPropDirect についても同様のラベル伝播手法だが、関連度は提案手法と同じく SBF と DBF の両方を利用して計算している。なお、表中の値は交差検定の平均値であり、精度 (Precision) と再現率 (Recall) から F 値 (F-measure) を計算すると多少のずれが生じる。

提案手法によるモデルが全ベースラインと比べて高い結果であり、ソース言語のリソースをラベル伝播と転移学習によってターゲット言語で利用する本アプローチは有効と言える。後半部の弱分類器を全て用いた OurMethodH2 と、信頼度 β_t でモデルを選定した OurMethodH3 を比較すると、前者の方が F 値は高かったが、モデル数は約 $(1/L)$ とコンパクトになっている。弱分類器数について、変動させた TrAdaBoost の結果から、本稿で設定した N よりも大きい数の方が高精度が期待できる。しかしながら、本手法はラベル伝播の工程で記憶領域を大きく割き、行列作成を並列処理させても学習時間を多大に消費する。近似して固有値を計算する手法の導入や、オンライン学習に基づく手法を取り入れていく必要があり、弱分類器数の増減とともに今後調査したい。

各ベースラインについて、BSLineTgt と BSLineSrcTgt の結果から、言語差が少ない SBF でも単純に学習データを足し合わせるだけでは精度向上は見られない。対して LbPropInDirect では、ソース言語の学習データが、同じくソース言語のラベルなしデータの擬似ラベルを通じて影響を及ぼしたと思われる。この点から、共通した観点で素性抽出した場合では、コーパスの言語に大きく依存しない可能性がある。また、ソース言語から直接ターゲット言語にラベルを伝播した LbPropDirect では、ほぼ全ての入力に対して正ラベル (+1) を返したために F 値自体は高い。提案手法でも再現率が高く、ソース言語で利用した学習データの正ラベルの割合がターゲット言語よりも大きいことに起因しているかもしれない。ラベル伝播時のパラメータや関連度の計算方法、使用する素性などと併せて詳しい分析と修正の余地がある。

4 まとめと今後の方針

本稿では、他言語リソース (辞書データやシソーラス、コーパス) を半教師あり学習 (ラベル伝播) と転移学習 (TrAdaBoost) を二段階で利用し、二値分類問題に定式化できる重要文抽出のシステム精度向上を図った。提案手法により、学習時間と記憶領域を消費するものの、ベースラインと比べて高い結果となった。

今後は近似による計算量の低減と並列・オンライン学習化などの実装面での改良や、弱分類器数や信頼度の影響などのパラメータについての調査、ソース言語とターゲット言語を入れ替えての実験やターゲット言語のリソース使用、より自動要約タスクに適した回帰モデルへの応用 [7] などを行っていく。

謝辞

NTCIR と DUC のテストコレクションがシステムの構築と評価に大きく貢献しました。心より感謝を申し上げます。

参考文献

[1] Xiaojin Zhu, Zoubin Ghahramani. Learning from Labeled and Unlabeled Data with Label Propagation. Technical Report CMU-CALD-02-107. 2002.

[2] Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux. Label Propagation and Quadratic Criterion. In Semi-Supervised Learning. MIT Press. pp.193-216. 2006.

[3] 神嶌敏弘. 転移学習. 人工知能学会誌. Vol.25. No.4. pp.572-580. 2010.

[4] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, Yong Yu. Boosting for Transfer Learning. In Proceedings of the 24th Annual International Conference on Machine Learning, pp.193-200. 2007.

[5] 神嶌敏弘, 濱崎雅弘, 赤穂昭太郎. 飼いならし — 飼育・野生混在データからの学習. 人工知能学会全国大会第 22 回論文集. 2D1-3. 2008.

[6] Yi Yao, Gianfranco Doretto. Boosting for transfer learning with multiple sources. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp.1855-1862. 2010.

[7] David Pardoe, Peter Stone. Boosting for Regression Transfer. In Proceedings of the 27th International Conference on Machine Learning. pp.863-870. 2010.

[8] Peilin Zhao, Steven C.H. Hoi. OTL: A Framework of Online Transfer Learning. In Proceedings of the 27th International Conference on Machine Learning. pp.1231-1238. 2010.

[9] Xiao Ling, Wenyuan Dai, Gui-Rong Xue, Qiang Yang, Yong Yu. Spectral Domain-Transfer Learning. In Proceeding of the 14th ACM International Conference on Knowledge Discovery and Data Mining. pp.488-496. 2008.

[10] Wenyuan Dai, Ou Jin, Gui-rong Xue, Qiang Yang, Yong Yu. EigenTransfer: A Unified Framework for Transfer Learning. In Proceedings of the 26th International Conference on Machine Learning. pp.193-200. 2009.

[11] 天野禎章, 横山晶一. 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, 言語処理学会第 16 回年次大会論文集. pp.47-50. 2010.

[12] Günes Erkan, Dragomir Radev. LexRank: graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research. Volume 22. Issue 1. pp.457-479. 2004.

[13] Zheng Wang, Yangqiu Song, Changshui Zhang. Knowledge Transfer on Hybrid Graph. In Proceedings of the 21st International Joint Conference on Artificial Intelligence. pp.1291-1296. 2009.

[14] Leo Breiman. Random Forests. Machine Learning. Volume 45. pp.5-32. 2001.

[15] Nello Cristianini, John Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press. 2000.