The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

言語・非言語情報の融合に基づく重要発言の推定

Predicting Important Utterance based on Fusing Verbal and Nonverbal Information

二瓶 芙巳雄[*1] 中野 有紀子[*2] 高瀬 裕[*2]
Nihei Fumio Nakano Yukiko Takase Yutaka

[*1] 成蹊大学大学院理工学研究科 Graduate School of Science and Technology, Seikei University
[*2] 成蹊大学理工学部 Faculty of Science and Technology, Seikei University

Automatic meeting summarization would reduce the cost of producing minutes during or after a meeting. With the goal of establishing a method for extractive meeting summarization, we propose a multimodal fusion model that identifies the important utterances that should be included in meeting extracts of group discussions. The proposed multimodal model fuses audio, visual, motion, and linguistic unimodal models that are trained by employing a convolutional neural network approach. The performance of the verbal and nonverbal fusion model presented an F-measure of 0.827. We also discuss the characteristics of verbal and nonverbal models and demonstrate that they complement each other.

1. はじめに

グループ議論は意思決定や新規アイデアの創出に役立つ. しかしながら, 議論で上がった意見を共有するため, また議論を振り返るためには, 議論の重要な点を議事録として要約する必要がある. 議論を自動で要約することができれば, 議事録作成のための作業を削減することができる.

自然言語処理の研究領域で発展した文書要約生成[Saggion 13, Spärck Jones 07]の技術は, 会議の要約のためにも使用されている. また会議はマルチモーダルインタラクションの側面を持つことから, テキスト情報以外にも, 会議の参加者の非言語情報も利用することができる. 多くの研究が要約に含めるべき重要発言の推定に, 文の長さや tf*idf などの言語情報に加え, 音声特徴量などの非言語情報も使用したモデルを提案している[Murray 06, Xie 09].

しかし以上の研究は, マルチモーダル情報の共起関係までを考慮したものではなかった. 例えば, 大きな声での発声は大きな身体動作と共起するといえ, そしてそれは重要発言を特徴づける行動と考えられる. さらには, グループ会議では個人の行動間の共起に加え, 複数人の行動間の共起関係も考慮する必要がある. 例えば, 他者からよく注視された発言はグループから重要だと認知されていると考えられる.

だがスピーチ, ジェスチャ, 表情, そして言語行動といった単一モダリティを表現する特徴量を定義した上で, 複数モダリティ間の共起関係を表す特徴量や, 複数の参加者にまたがる行動データの共起関係を表す特徴量を手作業で定義することは現実的でない. ここで深層学習には, 音声や画像からその特徴表現を自動で獲得できる[Golik 15, Pan 16]ことから, 単一モダリティの特徴表現だけでなく, 複数モダリティ間の特徴表現, さらには参加者間の特徴表現も自動で獲得することが期待できる.

以上の議論より本研究では, 複数参加者にまたがるマルチモーダル情報を深層学習によりフュージョンさせた, 高性能な重要発言推定モデルを提案する. さらには, 得られたフュージョンモデルの特性や性能向上の要因も調査する. そのためまず, 言語情報を使用するモデルや, スピーチや頭部動作に基づく非言語情報モデルを作成する. 非言語情報モデルは[Nihei 17]の研究成果を利用する. つぎに, 両者をフュージョンさせたモデルを作成し, 言語, 非言語, 言語・非言語フュージョンの 3 種類のモデルの推定性能を比較する. さらに, 本研究で提案する言語情報モデルと非言語情報モデルそれぞれがとらえる重要発言の特徴を調査し, さらに言語情報と非言語情報のフュージョンによる改善点についても議論する.

連絡先: 二瓶 芙巳雄, 成蹊大学大学院, 0422-37-3756, dd166201@cc.seikei.ac.jp

2. データ

本研究では MATRICS マルチモーダル議論コーパスを使用した[林 15]. これには, 4 人から形成される 9 つのグループによる, 計 16 回の対話が収録されている. 議論グループは意思決定型のタスクについて, 約 20 分間議論した.

収集データは, 会話参加者ごとのスピーチ, 顔映像, そして頭部動作データである. スピーチはヘッドセットマイクにより収録した. 顔映像は web カメラにより記録された. 頭部動作は議論参加者の後頭部に装着された加速度センサにより計測された. またスピーチから発話を認定した. 発話はアノテータがアノテーションツール上で, 音声波形を確認しながら認定した. 発話の区切りは, 所与の有声区間の前後に合計 300ms 以上の無音区間が生じた場合とした. そして各発話に対して, 発話内容を人手により書き起こした. 分析データの詳細は 3 章で説明する.

コーパス内の発話の重要度を評価するため, 5 名のアノテータが 16 の対話ビデオに含まれる 8,939 発話から, 議論の抜粋として採用すべきかという観点における重要発言を選択した. この 5 名のアノテータ間には相互に高い評価者間信頼性が確認されている(κ 係数 > 0.4). その後, 5 名のうち 3 名から重要発言と選択された発話を重要発言と定義した. 結果として, 3,789 発話が重要発言となった.

3. モデル

4 つの非言語情報ユニモーダルモデルと, 1 つの言語情報モデルを作成する(図 1). そしてこれらユニモーダルモデルの出力を統合し, マルチモーダルフュージョンモデルを作成する.

作成した全てのモデルは, 与えられた入力ベクトルが重要発言として選択されるべきか否かを判断するように訓練された.

3.1 非言語情報ユニモーダルモデル

先行研究[Nihei 17]において提案した, 3 次元データに対する畳み込み演算を行う軽量な CNN(図 1 左側)を, 非言語情報

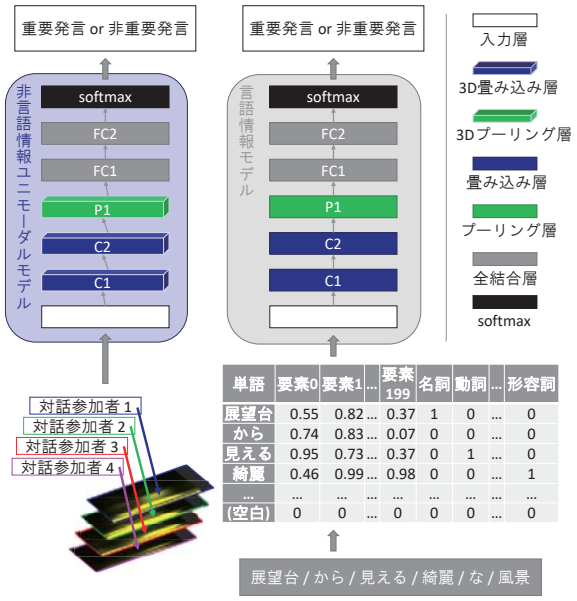

図 1 ユニモーダルモデルのネットワーク構造

ユニモーダルモデルとして使用する. このネットワークは, 2 層の 3 次元畳み込み層と 3 次元プーリング層, 2 層の全結合層, そして分類のための softmax 層からなる. このネットワークは議論参加者を畳み込み演算の要素として扱うこと, また膨大な量のデータが使用できない場合に多層ネットワークに引き起こされる過学習を回避することをねらったものである.

非言語情報ユニモーダルモデルのネットワーク構成を表 1 に示す. 収集されたセンサデータのサンプリング周波数がモダリティによって異なるため, 入力ベクトルのサイズはモダリティにより異なる. また, ネットワークへの入力となる発話の時間長をすべて 15 秒に統一するために, 時間長が短い発話には空白のデータを追加した. したがって, 入力ベクトルの幅はサンプリング周波数×15 になる. 全ての非言語情報ユニモーダルモデルは, SGD(AdaDelta, ミニバッチサイズ 32)で訓練した. エポック回数は 30 であり, 全ての全結合層と畳み込み層における活性化関数は ReLu を用いた. 図 1 における全結合層のユニット数は FC1, FC2 いずれも 128 である. 畳み込みカーネルの数は, C1, C2 ともに 32 である. P1 層と FC1 層間のドロップアウト率は 0.25 であり, FC1-FC2 間, また FC2-softmax 間は 0.5 である.

モデルの入力ベクトルとして使用するモダリティは, 頭部動作スペクトログラム, スピーチスペクトログラム, スピーチインテンシティ, 頭部姿勢である. 以降では各入力ベクトルの表現とネットワークの構成ついて詳細を説明する.

(1) 頭部動作スペクトログラムモデル (HS model)

加速度センサは各議論参加者の後頭部に取り付けられ, 30fps で 3 軸の角速度を出力する. 角速度を以下の式(1)に適用することで, 各議論参加者の頭部合成角速度を算出した. ここでx_i, y_i, z_iはそれぞれ, iフレーム目におけるx, y, z軸の角速度である.

$$HMA_i = \sqrt{x_i^2 + y_i^2 + z_i^2} \quad \cdots (1)$$

時系列の頭部合成角速度データを, フーリエ変換を用いて, 時間, 周波数成分, そして振幅の大きさを表すスペクトログラムとして表現した. フーリエ変換のウィンドウ幅は 30 フレーム, スライド幅は 1 フレームとした. スペクトログラムは各発話区間に対して, 4 人の各議論参加者について作成される. 従って表 1 に示すように, 4 人の議論参加者から得られる頭部動作データは,

表 1 各ユニモーダルモデルに対する, 入力・畳み込みカーネル・プーリングフィルタそれぞれのサイズ

ユニモーダルモデル	入力サイズ	畳み込みカーネルサイズ	プーリングフィルタサイズ
HS model	450, 15, 4, 1	3, 3, 4	2, 2, 1
SS model	750, 32, 4, 1	5, 3, 4	
SI model	1500, 1, 4, 1	10, 1, 4	
HP model	450, 3, 4, 2	3, 3, 4	
V model	48, 219, 1	48, 1	2, 1

450(発話時間長=30fps× 15sec)×15(周波数分解能)×4(議論参加者数)×1(チャネル数)のサイズとなる. これをネットワークに対する入力とする.

畳み込み層のカーネルサイズは 3×3×4 であり, プーリングサイズは 2×2×1 である.

(2) スピーチスペクトログラムモデル (SS model)

各議論参加者に装着したヘッドセットマイクからサンプリング周波数 44.1kHz の音声を記録し, そこからスペクトログラムを作成した. フーリエ変換のウィンドウ幅は約 1.5 秒($2^{16} = 65{,}536$フレーム), スライド幅は 1 フレームとした. 従って, 観測できる最大周波数は約 32kHz となる. またフーリエ変換したのちのサンプルを 50fps になるようにダウンサンプルした. 加えて最大で約 32kHz 観測できる周波数成分を 32 区間に分割した. そして, 分割された各区間における周波数の強度の総和をデータポイントとした. 頭部動作モデルと同様に, スペクトログラムは各発話区間に対して, 4 人の各議論参加者について作成される. 従って 750(発話時間長=50fps× 15sec)×32(周波数分解能)×4(議論参加者数)×1(チャネル数)のサイズとなる. 畳み込み演算のためのカーネルサイズとプーリングサイズは表 1 に示す.

(3) スピーチインテンシティモデル (SI model)

各発話について, スピーチインテンシティを音声分析ツール [1] により 100fps で計測し, 1500(発話時間長=100fps×15sec)×1(次元数)×4(議論参加者数)×1(チャネル数)のサイズのデータを作成した. ネットワークの詳細な設定は表 1 に示す.

(4) 頭部姿勢モデル (HP model)

議論参加者の顔面を記録した映像に対して, 画像処理による顔検出器 [2] を適用することにより, x,y,z 軸における頭部位置と回転角を 30fps で計測した. それぞれは, 450(発話時間長=30fps×15sec)×3(x,y,z それぞれ)×4(議論参加者数)×2(チャネル数, 位置と回転角)のデータに変形された. 詳細なネットワーク設定は表 1 に示す.

3.2 言語情報モデル (V model)

言語情報モデルは各発話区間の発話内容を入力とするため, ネットワークへの入力は一人の 1 発話の言語情報となる. そのため, ここで作成するネットワークは, 非言語情報ユニモーダルモデルで行った 3 次元的な畳み込み演算ではなく, 2 次元的な畳み込み演算を行う CNN となる. ネットワークの構造は図 1 に示すように, 非言語情報ユニモーダルモデルと同様であるが, 3 次元畳み込み層と 3 次元プーリング層がそれぞれ, 2 次元畳み込み層と2次元プーリング層に置き換わる. 訓練に使用したハイパーパラメタは全て, 非言語情報ユニモーダルモデルと同一である.

各発話区間の発話内容を単語に分割し, 各単語を skip-gram モデルによりベクトル化した. 形態素解析器には mecab[3] を使用

[1] Praat: http://www.fon.hum.uva.nl/praat/
[2] FaceAPI: https://www.seeingmachines.com/
[3] Mecab: http://taku910.github.io/mecab/

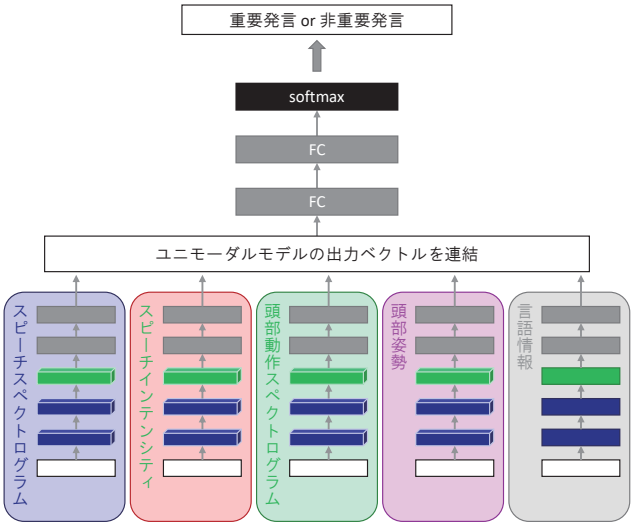

図 2　言語・非言語フュージョンモデル

し，新語対応辞書である NEologd[4] を使用した．Skip-gram モデルの訓練には wikipedia の記事に含まれる文章を用いた．訓練されたモデルは単語を 200 次元のベクトルで表現する．加えて，形態素解析では，各単語に 19 種類の品詞のいずれかが付与されるが，これを 19 次元のベクトルとして表現した．この 2 つを統合し，各単語について 219 次元のベクトルを作成した．非言語情報と同様の考えから，発話内の単語数が少ない発話には，単語数が 48 になるように，空白の単語ベクトルを追加した．48 は発話内容における最大単語数である．したがって，入力データのサイズは，48（最大単語数）×219（単語ベクトル 200+品詞ベクトル 19）×1（チャネル数）となる．詳細なネットワーク設定は表 1 に示す．

3.3 マルチモーダルフュージョンモデル

ユニモーダルモデルを統合し，非言語フュージョンモデル (NV fusion)と，言語・非言語フュージョンモデル(V-NV fusion，図 2 参照) の，2 つのフュージョンモデルを作成した．NV fusion はすべての非言語情報ユニモーダルモデルを統合したものであり，V-NV fusion はすべての非言語情報ユニモーダルモデルと言語情報モデルを統合したものである．ユニモーダルモデルを統合するにあたり，それぞれの softmax 層を破棄し，softmax 層に接続していた全結合層から得られる出力ベクトルを連結させ，フュージョンモデルの入力ベクトルとした．従って，NV fusion と V-NV fusion の入力ベクトルの次元数はそれぞれ，512（=128×4），640（=128×5）である．訓練に使用したハイパーパラメタはドロップアウト率を除き，非言語情報ユニモーダルモデルと同一である．マルチモーダルフュージョンモデルにおいて，ドロップアウトは使用されない．

4. モデル性能の評価

前節で提案したモデルの性能を，Leave-One-Group-Out 交差検証法により評価する．この交差検証法は，訓練に使用しないグループのデータを用いてモデルを評価するため，未知なデータが供給された時のモデル性能を最も適切に評価できる．訓練データはアンダーサンプリングにより正例数と負例数を一致させた．

図 3 に，F 値によるモデルの評価結果を示す．V-NV fusion が最良で，その性能は 0.827 であった．言語情報と非言語情報を

[4] NEologd: https://github.com/neologd/mecab-ipadic-neologd

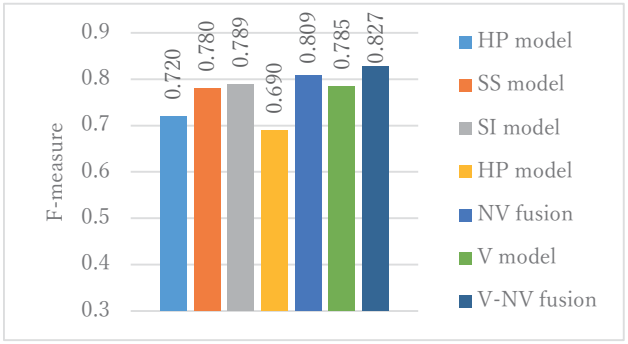

図 3　モデル性能

フュージョンすることにより，いずれのユニモーダルモデルよりも高い性能を示した．従来の機械学習を用いた研究でも報告されたように，深層学習によるアプローチにおいてもマルチモーダルフュージョンは有効であった．次に優れていたモデルは，全ての非言語ユニモーダルモデルを統合した NV fusion であった．ユニモーダルモデルの中では，SI モデルと SS モデル，そして V モデルが優れていることから，音声情報と言語情報を用いたモデルは高性能であった．

5. 議論

本章では，V モデルと NV fusion モデル，そして V-NV fusion モデルの特性について考察する．

5.1 言語情報モデルの特性

言語情報モデルが非言語情報モデルよりも高性能に重要発言を推定できる場合を観察することで，言語情報でのみ捉えられる重要発言の特徴を調査する．

まず，コーパス内で 10 回以上正例として出現する品詞の組み合わせのうち，V モデルが NV fusion モデルよりも多く正例として推定できた組み合わせを数え上げた．その結果を表 2 に示す．言語情報モデルが非言語情報モデルよりも優れた推定を発揮する品詞の組み合わせには，名詞と助詞が必ず含まれており，かつ動詞など，述語と判断できる語も多くの場合で含まれていた．そしてそれらの発話は，ある名詞に対する議論や意思決定，またアイデアの提案であることが多い．

さらに，コーパス内で 20 回以上負例として出現した品詞の組み合わせのうち，V モデルが NV fusion モデルよりも多く負例として推定できた組み合わせについても確認した．その結果，単語数は 1 から 4 であり，さらにフィラーや感嘆詞が含まれる品詞の組み合わせが多いことが特徴的である．そして典型例は，“ああそうですね”，“うん多分”，“ありですね”，“良さげですけど”といった，相槌や確認，また独り言のような，命題的情報を伝達

表 2　V モデルが NV fusion モデルより正例を多く推定できた品詞の組み合わせ

単語数	品詞の組	典型例
3	助詞/動詞/名詞	築地行って
	助詞*2/名詞	サイドメニューとかは
	助詞/助動詞/名詞	焼きそばとかで
	助詞/接続詞/名詞	じゃ東京で
4	助詞*2/名詞*2	パンケーキか唐揚げか
	助詞*2/動詞/名詞	成田着いてから
	助詞/接続詞/名詞*2	じゃあそこで昼
	助詞/動詞/副詞/名詞	まあスカイツリー見て
5	助詞/助動詞/動詞/名詞*2	皇居行くんですか
7	助詞*3/助動詞/名詞*3	上野と秋葉原どっちですかね

する意図のない発話であった．従って言語情報モデルは非言語情報モデルより，命題的情報を伝えない発話を非重要発言としてより多く，より正しく推定することができる．

5.2 非言語情報モデルの特性

非言語情報モデルが言語情報モデルよりも高性能に重要発言を推定できる場合を観察することで，非言語情報モデルがとらえる重要発言の特徴を調査する．

まず，NV fusion モデルが正例の推定に成功した一方，V モデルは推定に失敗した 450 発話を抽出した．そしてそれら発話における，スピーチスペクトログラムと頭部動作スペクトログラムの平均画像を作成した（図 4）．この状況下において，発話者のスピーチスペクトログラムにはエネルギーが確認できるが，その他の参加者のスピーチスペクトログラムにはエネルギーが確認できない．すなわち，ある参加者が発声するとき，その他の参加者は発声していないことが示唆される．頭部動作スペクトログラムでは，発話者のものはエネルギーが強いが，その他の参加者のものは比較して弱い．以上より NV fusion モデルは，他者から静かに傾聴され，注目された発言を重要発言として正しく検出することが示唆された．

一方で NV fusion モデルが負例の推定に成功するが，V モデルは推定に失敗した 503 発話で同様の観察を行った結果，発話者とその他の参加者それぞれのスピーチスペクトログラムにはいずれも強いエネルギーが生じていた．また同様の傾向が頭部動作スペクトログラムにも確認できたことから，NV fusion モデルは全員が発声し，かつ活発に動作する場合に，非重要発言を正しく除外することが可能といえる．

5.3 言語情報と非言語情報のフュージョンによる改善

V-NV fusion モデルによる推定結果から得られる混合行列を，V モデル，また NV fusion モデルと比較することで，主な性能改善点を明らかにする．表 3 に，V モデル，NV fusion モデルそれぞれの混合行列を基準としたときの，V-NV fusion の混合行列の改善割合を示す．V モデルと NV fusion モデルどちらを基準としたときにも，FP が大きく減少し，かつ TN が上昇している．すなわち，非言語情報と言語情報をフュージョンさせることで，非重要発言を誤って重要発言と推定することによる重要発言検出の適合率の減少を防ぎ，かつ非重要発言を正しく排除することが増えるようになることが確認できた．

6. おわりに

本研究では畳み込みニューラルネットワークを用いて，グループ議論における重要発言の推定モデルを提案した．モデルは形態素解析により得られた単語と品詞情報を用いた言語情報モデルと，これを先行研究[Nihei 17]で提案した非言語情報フュージョンモデルと統合した言語・非言語フュージョンモデルを作成した．その結果，言語・非言語フュージョンモデルは全てのモデルの性能を上回り，かつ F 値 0.827 で重要発言を検出可能な高性能なモデルが作成できた．また言語情報モデルの特性を分析した結果，名詞，助詞，動詞といった文の基本的な構成素を含む発話に対しては，非言語情報モデルよりも高性能で推定でき，また非言語情報モデルは他者から静かに傾聴され，注目された発言を言語情報モデルより高性能に推定できること

表 3　V-NV fusion モデルの改善割合

基準	TP	TN	FP	FN
NV fusion モデル	0.34%	3.67%	-16.83%	-1.86%
V モデル	5.07%	4.08%	-18.34%	-21.12%

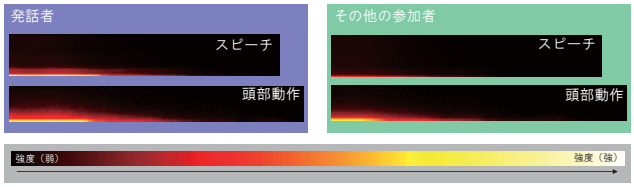

図 4　スピーチと頭部動作の活動量

が確認された．また言語情報と非言語情報をフュージョンすることで，誤推定が減少する効果があることが示唆された．

将来的には本研究の最終目標である，議論の自動要約システムを実現させるため，提案した重要発言推定モデルを用いた要約生成アルゴリズムを検討する．さらに有用な要約が生成できたかを検証するため，生成された要約の評価実験を実施する．

謝辞

本研究は，科学技術振興機構(JST) 戦略的創造研究推進事業(CREST)「実践知能アプリケーション構築フレームワーク PRINTEPS の開発と社会実践」(JPMJCR14E3)，および理化学研究所革新知能統合研究センターの支援を受けたものである．

参考文献

[Golik 15] Golik, P., Tüske, Z., Schlüter, R., and Ney, H.: Convolutional neural networks for acoustic modeling of raw time signal in LVCSR, In *INTERSPEECH-2015* (pp. 26–30). Dresden, Germany, (2015)

[Murray 06] Murray, G., Renals, S., Carletta, J., and Moore, J.: Incorporating Speaker and Discourse Features into Speech Summarization, In *Proceedings of NAACL HLT* (pp. 367–374). Stroudsburg, PA, USA: Association for Computational Linguistics, (2006)

[Nihei 17] Nihei, F., Nakano, Y. I., and Takase, Y.: Predicting Meeting Extracts in Group Discussions Using Multimodal Convolutional Neural Networks, In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 421–425). New York, NY, USA: ACM, (2017)

[Pan 16] Pan, J., Sayrol, E., Giro-I-Nieto, X., McGuinness, K., and O'Connor, N. E.: Shallow and Deep Convolutional Networks for Saliency Prediction, In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 598–606). (2016)

[Saggion 13] Saggion, H., and Poibeau, T.: Automatic Text Summarization: Past, Present and Future, In T. Poibeau, H. Saggion, J. Piskorski, & R. Yangarber (Eds.), *Multi-source, Multilingual Information Extraction and Summarization* (pp. 3–21). Berlin, Heidelberg: Springer Berlin Heidelberg, (2013)

[Spärck 07] Spärck Jones, K.: Automatic Summarising: The State of the Art, Inf. Process. Manage., Vol. 43, No. 6, pp. 1449–1481. (2007)

[Xie 09] Xie, S., Hakkani-Tur, D., Favre, B., and Liu, Y.: Integrating prosodic features in extractive meeting summarization, In *IEEE Workshop on Speech Recognition and Understanding (ASRU)* (pp. 387–391). (2009)

[林 15] 林佑樹，二瓶芙巳雄，中野有紀子，黄宏軒，and 岡田将吾: グループディスカッションコーパスの構築および性格特性との関連性の分析，情報処理学会論文誌，Vol. 56, No. 4, pp. 1217–1227. (2015)