A multitude of linguistically-rich features for authorship attribution

Ludovic Tanguy, Assaf Urieli, Basilio Calderone,
Nabil Hathout, and Franck Sajous

CLLE-ERSS: CNRS & University of Toulouse, France

PAN 2011 Workshop – Authorship Attribution - CLEF

- Who we are
  - A small NLP team in a linguistics lab
  - Fields ranging from morphology to discourse
  - More and more involved in document classification
- Our Motivations
  - Compete in a challenge where several innovative linguistic features can be used
  - Assess the usefulness of richer features
- Our Method
  - Annotation with many features, most of them being linguistically-rich
  - Maximum Entropy classifier and some rule-based learners

Linguistic Features

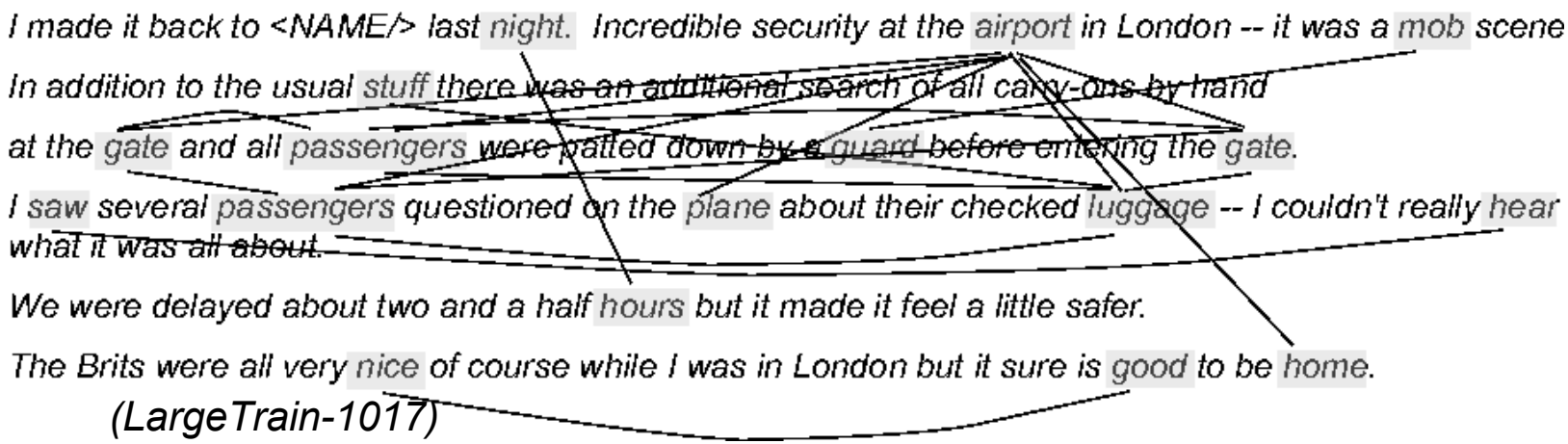
- What are «rich» features?
  - They make use of external language knowledge
  - They require more complex operations than word lookup
- Examples
  - Morphology: suffix frequency (CELEX database)
  - Syntax: sentence complexity (Stanford parser)
  - Semantics:
    - *Ambiguity and specificity (WordNet)*
    - *Cohesion (semantic links from the Distributional Memory database)*
  - Adhoc features
    - *Spelling errors, openings and closings, etc.*

- Morphological complexity
  - Based on suffixes extracted from the CELEX morphological database
  - High frequency of suffixed words
    - *<NAME/>, Attached is a clean document for **execution**. If in **agreement**, please sign two **originals** and forward same to my **attention** for final signature. I will return a **fully** executed **agreement** for your records. Do not hesitate to give me a call should you have any **questions** regarding the enclosed.*
      Best regards,
      (SmallTrain-2249)
  - Low frequency of suffixed words
    - *Suz, I say lets do it! and so does <NAME/>. I will make Rotel dip and other stuff too. I think it will be fun - and maybe we can carry the party to the hood after! Keep me posted on how your day is going. I kind of hope you get to go today to see your fam.*
      K.
      (SmallTrain-1358)
  - Also specific suffixes (-*ous*, -*ing*, etc.)

- Syntactic complexity
  - Based on Stanford dependency parser
    - *Syntactic tree depth*
    - *Distance between syntactically dependent words*
  - High complexity (avg distance 3.6, avg depth 8.5) :
    - *unfortunate...**but you also don't want to go getting yourself attached to someone whom you ultimately don't have enough in common with to sustain the kind of relationship you're looking for**. off the soapbox.... i'm going to the grocery store (forgot some things), the dry cleaners, running and finishing up laundry detail...so that will take up a bit of time. (SmallTrain-2944)*
  - Low complexity (avg distance 2.7, avg depth 2.7)
    - *<NAME/>,Seattle was sweet this weekend. I went and saw <NAME/> at the Breakroom...what did you think of Husky Stadium? Woohoo! Man, Thursday...whoa...and think, I went out after that...whoa...but it was my birthday...sorry for calling late. Are you doing anything cool this weekend? Motorcycle dirt track races are on Saturday night at Portland Speedway…I am stoked. Plus the first Husky football game is this weekend in Seattle against Michigan! How are other things going? Hopefully well. Later, <NAME/> (SmallTrain-623)*

- Semantic specificity and ambiguity
  - Number of WordNet synsets per word and average depth of synsets in the hierarchy (*specific* *generic*)
  - High specificity
    - *Hey <NAME/>,*
      I've done some research on the actuals that you make reference to (Vectren). <NAME/>'s sale with Heartland Steel is at the interconnect between Midwestern Gas Transmission and Vectren (formerly know as Indiana Gas). The actual volumes that you are reporting and consider to be your monthly actuals are volumes that I believe are behind Vectren's city gate (which means that you more than likely have an imbalance on Vectren's system). This bears checking with Vectren, regarding an imbalance behind their gate. You should be receiving some type of statement or invoice from Vectren. Per the contract, <NAME/> uses the Midwestern Gas Transmission (pipeline statement) to actualize our monthly invoices to you. I've attempted to draw a diagram for you to make it as clear as I can.
      Let's talk!
      (SmallTrain-929)
  - Low specificity (generic vocabulary)
    - *I believe that we did have some activity on Blue Dolphin, but it was done by the Wellhead group. You should send the Vol Mgmt people to <NAME/> Smith.*
      (SmallTrain-2579)

- Cohesion
  - Based on Distributional Memory (Baroni & Lenci 2010)
    - *Words are related if they appear in the same syntactic contexts in a reference corpus*
    - *Measure rate of related word pairs in the message*
  - High cohesion



(LargeTrain-1017)

  - Low cohesion (no links)
    - *We are OUT of the pool. I want my money back. Prentice, please get your stuff out of my apartment. You can have the cats.*
      Love,
      <NAME/>
      (LargeTrain-2285)

- Ad hoc features:
  - Sample opening patterns (22):
    - *<NAME/>*
      Dear <NAME/>
      Hello
      <NAME/>,
      Hi,
      Hey <NAME/>,
      Hello <NAME/>
      Hi <NAME/>,
  - Sample closing patterns (44):
    - *thanks,\n*
      Thanx
      <NAME/>\n
      thank you,\n<NAME/>
      Love,\n<NAME/>
      thanks,\n<NAME/>.
      best,\n<NAME/>
      thanks!\n<NAME/>
      Thanks.

- Additional « poorer » features we used:
  - Character trigrams
  - Word frequencies (Bag of words)
  - Punctuation marks
  - Part-of-speech tags unigrams & trigrams
  - Named entities
  - Length of words, messages, lines
  - Use of blank lines
  - Contractions
  - US/UK vocabulary
  - etc.

Machine Learning

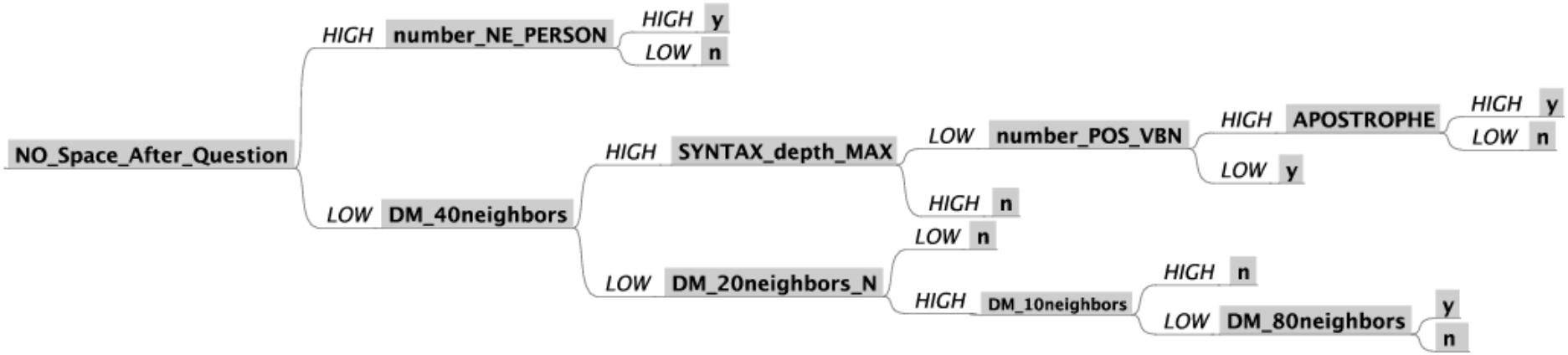
- Two objectives:
  - Manage a large number of very different features
  - Get some feedback from the models
    - *At least the relative contribution of individual features*
    - *If possible, some clues about each author's most discriminant features*
- Two methods:
  - Author identification (success): Maximum entropy classifier
  - Author verification (failure): C4.5 Decision tree and RIPPER

- More details about Maximum Entropy
  - OpenNLP Maxent (http://incubator.apache.org/opennlp)
  - CSVLearner
    - *Homegrown software for normalization, training and evaluation (https://github.com/urieli/csvLearner)*
  - No preliminary feature selection
    - *Except for character trigrams, only the most frequent 10,000 (freq>12)*
  - Numeric features (i.e. distances, relative frequencies etc.) normalised based on max values in the training data
    - *Some groups of features (e.g. PosTag trigrams) normalised based on max value in entire group*
  - Nominal and boolean features used as such

- Dealing with unknown authors
  - MaxEnt gives a probability for each author
  - Remarkable correlation between low MaxEnt decision probabilities and errors/unknown authors.
  - If top author probability < threshold, set output to « unknown »
    - *Two runs submitted with different thresholds*

Set	Threshold	Macro Prec	Macro Recall	Macro F1	Micro Prec	Micro Recall	Micro F1
SmallTest+	66%	73.7	**16.1**	**19.3**	82.4	**45.7**	**58.8**
SmallTest+	95%	**95.5**	6.8	10.7	**96.6**	18.0	30.3
LargeTest+	40%	68.8	**26.7**	**32.1**	77.9	**47.1**	**58.7**
LargeTest+	75%	**80.6**	14.8	20.8	**92.4**	29.9	45.1

- Rule-based learning for author verification
  - Verify1 (C4.5):



  - Verify2 (RIPPER):
    - *1. if DM90neighbors ≥ 0.00493 and DM80neighbors ≤ 9 and APOSTROPHE = 0 then Y*
    - *2. if DM20neighbors ≥ 0.0173 and COLON ≥ 0.0090 and DM10neighbors ≤ 28 then Y*
    - *3. otherwise N*

- Rewriting history… If we had used the maximum entropy classifier for author verification
  - Adding 100 random messages from training sets
  - Results on the test sets:

Task	Method	Precision	Recall	F-score
Verify1	Decision tree (submitted)	0.09	0.33	0.143
	Max.Ent	***0.33***	***0.66***	***0.444***
Verify2	RIPPER (submitted)	0.1	0.2	0.133
	Max.Ent	***1***	***0.40***	***0.571***
Verify3	RIPPER (submitted)	0.08	0.25	0.125
	Max.Ent	***0.25***	***0.25***	***0.25***

- At least double the performances!

A look at the models

- Rule-based learners
  - A good variety of features were selected by both methods on the three tasks
  - Very few low-level features emerged
  - … But of course a very poor performance!
- Maximum entropy: assessing features
  - Compare overall results with and without features sets
  - Have a look at the trained model
    - *Author/feature coefficients*

- With and without richer features
  - Comparison between different sets of features
    - *Training: SmallTrain. Evaluation: SmallTest.*

Features	Total Accuracy	Avg. Precision	Avg. Recall	Avg. F1
Rich	61.01	40.13	35.11	36.17
Poor	68.08	45.91	37.62	38.03
All	70.30	58.28	41.20	43.39
All - Poor	**+2.22**	**+12.37**	**+3.58**	**+5.36**

  - Conclusion: small but significant improvement over poorer features, but these are still needed

- Per author, extracting the most distinctive features from the MaxEnt model
  - Apply the trained model to all of the author's messages in the training set.
  - For each feature, sum up the weight that was attributed to the current author on each message

- Total weight distribution in the trained MaxEnt model:

  SmallTrain dataset

  - *Character 3grams:* *54%*
  - *Word unigrams:* *11%*
  - *POS 3grams:* *10%*
  - *POS unigrams:* *4%*
  - *Rich features:* *22%*
    - Morphology: 4%
    - Syntax: 6%
    - Semantics: 5.5%
    - Others: 6%

- Authors characteristics
  - Focus on target authors of the Verify tasks
- Distinctive features as given by MaxEnt weights
  - Author1:
    - *blank lines, determiners, number of sentences, high ambiguity nouns, no signature…*
  - Author2:
    - *<NAME/> elements, suffixes, uppercase words…*
  - Author3:
    - *blank lines, full stops, number of lines, number of sentences, syntactic complexity…*

- Human intuition on authors' distinctive features
  - Author1:
    - *Interrogative sentences without a « ? » (5/9 interrogative)*
    - *Automatically generated e-mails (17/42 messages)*
      - *The report named XXX, published as of YYY is now available…*
  - Author2:
    - *Short sentences and short messages*
    - *Shifts in person (from « I » to « we »)*
      - 10/50 messages
      - ***I** have a few thoughts on the offsite. **I** think **we** could have a theme of restructuring and change. **We** would have to make sure it is forward looking and upbeat in that **we** have learned a lot that will make **us** better in the future.*

- Author3:
  - *Lots of modalising verbs with a 1st person subject*
    - 41/105 verb occurrences
    - *Know, hope, doubt, mind, feel, like, think, enjoy, guess*, etc.
  - *Combinations of « Let me know » and « if/how/wh.. »*
    - 10/37 messages
    - *If you have any problems, let me know.*
    - *Please let me know if you know where <NAME/> is.*
    - *Let me know if this interferes with any plans.*

- Remarks
  - Most of the striking characteristics have not been measured (yet)
  - The others do not stand out in the trained model

Conclusions

- Good results on our first try at the task

- Still not sure about which is our main asset:
  - Linguistically rich features
  - MaxEnt classifier
  - Beginners' luck

- If the rich features are effectively a good thing
  - They still need a support from raw features
    - *This may be one of the explanation why the rule-based schemes failed*

- Further work
  - Statistical analyses to examine features
    - *Distribution, correlation, selection*
  - Other data sets and tasks
  - Still more features to design and use
    - *Sentence structures*
- Some thoughts about the task
  - Many rich features are in fact related to specific genres:
    - *formal mail to customers,*
    - *informal mail to family/friend,*
    - *short request/order to subordinates,*
    - *simple reply,*
    - *love letter,*
    - *etc.*
  - Could an author's « style » be defined as « features per genre » ?