Probabilistic Bag-of-Hyperlinks Model for Entity Linking

Octavian Ganea   Marina Ganea   Aurelien Lucchi
Carsten Eickhoff   Thomas Hofmann

Department of Computer Science
ETH Zürich, Switzerland

April 15, 2016
Today’s Menu

- Clean and effective solution for entity disambiguation
- State-of-the-art method (on Gerbil platform [Usbeck et al. 2015])
- Expands on the seminal works of
  - [Ferragina et al. 2010]
  - [Hoffart et al. 2011]
  - [Han et al. 2011]
  - [Ratinov et al. 2011]
Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.
Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.

- Not our focus
- Assume mentions are given as input in all experiments
- Typically found using a NER system
Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.
Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.
Local Disambiguation

Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.
Local Disambiguation: Mention - Entity Compatibility

Deutscher Bundestag  
Germany
Local Disambiguation: Mention - Entity Compatibility

Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.
Local Disambiguation: Mention - Entity Compatibility

- Commonness (see e.g. [Milne et al. 2008], [Ferragina et al. 2010])
- **Mention-Entity baseline:** \( e^*_i = \arg \max_{e \in \mathcal{E}} P(e_i|m_i) \)
- Estimated from Wikipedia statistics

\[
P(e|m) \approx \frac{\# \text{ links with } m \text{ that point to } e}{\# \text{ links with anchor } m}
\]
Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.

Table 1: Accuracy gains of individual PBoH components.
Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.
Local Disambiguation: Surrounding Context

Thomas Müller, the *midfielder* of *Germany*, scored one *goal* against Brazil in the final of the cup.
Local Disambiguation: Surrounding Context

Thomas Müller, the *midfielder* of Germany, scored one *goal* against Brazil in the final of the cup.

- Context = text window around mention (say of size 100)
- Similar context models: [Ratinov et al. 2011], [Blanco et al. 2015]
- Bag-of-words context model:

\[
P(c|e) = \prod_{w \in c} P(w|e) \quad P(w|e) \approx \frac{\text{#times } w \text{ in context of link to } e}{\text{#words surrounding links to } e}
\]

- Smoothing: absolute discounting with backoff interpolation
Local Disambiguation: Surrounding Context

Thomas Müller, the *midfielder* of Germany, scored one *goal* against Brazil in the final of the cup.

- Bayes’ rule (w/ conditional indepence assumption: \( c \perp\!\!\!\!\!\!\!\perp m \mid e \))

\[
P(e|m, c) \propto P(e|m)P(c|e) = P(e|m) \prod_{w \in c} P(w|e) \quad \text{(Local Context)}
\]

- Incremental accuracy:

<table>
<thead>
<tr>
<th></th>
<th>CoNLL test A</th>
<th>CoNLL test B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baselines</strong></td>
<td>R@MI</td>
<td>R@MA</td>
</tr>
<tr>
<td>Mention-Entity</td>
<td>69.73</td>
<td>69.30</td>
</tr>
<tr>
<td><strong>Local Context</strong></td>
<td>82.50</td>
<td>81.56</td>
</tr>
</tbody>
</table>
Joint Disambiguation

Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.
Joint Disambiguation

Thomas Müller

Germany

final

cup

\[ \rho_{e_j,m_j} = \log P(e|m) \]
Joint Disambiguation

Thomas Müller

Germany

final

cup

\[ \lambda_{e_i, e_j} = \log \left( \frac{P(e_i, e_j)}{P(e_i)P(e_j)} \right) \]

\[ \rho_{e_j, m_j} = \log P(e|m) \]
Joint Disambiguation

\[ \lambda_{e_i, e_j} = \log \left( \frac{P(e_i, e_j)}{P(e_i)P(e_j)} \right) \]

\[ \rho_{e_j, m_j} = \log P(e|m) \]
Joint Disambiguation

- Challenge: should not disambiguate single mentions in isolation

- [Han et al. 2011], [Ratinov et al. 2011], [Guo et al. 2014]

- We want to leverage entity-entity co-linking statistics:

\[ P(e, e') \approx \frac{\text{#articles have links to } e \text{ and } e'}{\text{#articles}} \]

- Log-linear model:

\[
P(e|m) = P(e_1 \ldots e_n|m_1 \ldots m_n) = \frac{1}{Z(m)} \exp \left[ \sum_{1 \leq i \leq n} \rho_{m_i,e_i} + \sum_{1 \leq i < j \leq n} \lambda_{e_i,e_j} \right]
\]
Markov Network and Factor Graph

- Probabilistic model: pairwise Markov Random Field

\[
P(e|m) \propto P(e)P(m|e) = P(e_1, \ldots, e_n) \cdot \prod_{i=1}^{n} P(m_i|e_i)
\]

⇒ prior \cdot likelihood

- Markov assumption: mentions independent given entities.

- Challenge: estimate \(P(e) = P(e_1, \ldots, e_n)\)
Markov Network and Factor Graph

- Plug-in estimators:
  
  \[ \rho_{m,e} = \log(P(e|m)), \quad \forall e, m \]
  \[ \lambda_{e,e'} = \log \left( \frac{P(e,e')}{{P(e)P(e')}} \right), \quad \forall e, e' \]

- Related to Bethe free energy (see paper)

- Exact for directed acyclic graphs. Approximation for loopy graphs.

<table>
<thead>
<tr>
<th>Baselines</th>
<th>CoNLL test A</th>
<th>CoNLL test B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@MI</td>
<td>R@MA</td>
</tr>
<tr>
<td>Mention-Entity</td>
<td>69.73</td>
<td>69.30</td>
</tr>
<tr>
<td>Local Context</td>
<td>82.50</td>
<td>81.56</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>69.77</td>
<td>69.95</td>
</tr>
</tbody>
</table>
Calibrated Model

Uncalibrated model:

$$P(e|m) = \frac{1}{Z(m)} \exp \left[ \sum_{1 \leq i \leq n} \rho_{m_i,e_i} + \sum_{1 \leq i < j \leq n} \lambda_{e_i,e_j} \right]$$

- The likelihood scales with $n$.
- The prior scales with $\frac{n(n-1)}{2}$. 

Based on a combinatorial argument (details in the paper).
Calibrated Model

Uncalibrated model:

\[ P(e|m) = \frac{1}{Z(m)} \exp \left[ \sum_{1 \leq i \leq n} \rho_{m_i,e_i} + \sum_{1 \leq i < j \leq n} \lambda_{e_i,e_j} \right] \]

Calibrated model:

\[ P(e|m) = \frac{1}{Z(m)} \exp \left[ \sum_{1 \leq i \leq n} \rho_{m_i,e_i} + \frac{2}{n-1} \cdot \sum_{1 \leq i < j \leq n} \lambda_{e_i,e_j} \right] \]

- Based on a combinatorial argument (details in the paper)
Calibrated Model

\[ P(e|m) = \frac{1}{Z(m)} \exp \left[ \sum_{1 \leq i \leq n} \rho_{m,i,e_i} + \frac{2}{n-1} \cdot \sum_{1 \leq i < j \leq n} \lambda_{e_i,e_j} \right] \]

where

\[ \rho_{m,e} = \log(P(e|m)), \quad \forall e, m \]

\[ \lambda_{e,e'} = \log\left( \frac{P(e,e')}{P(e)P(e')} \right), \quad \forall e, e' \]

<table>
<thead>
<tr>
<th>Baselines</th>
<th>CoNLL test A</th>
<th>CoNLL test B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@MI</td>
<td>R@MA</td>
</tr>
<tr>
<td>Mention-Entity</td>
<td>69.73</td>
<td>69.30</td>
</tr>
<tr>
<td>Local Context</td>
<td>82.50</td>
<td>81.56</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>69.77</td>
<td>69.95</td>
</tr>
<tr>
<td>Calibrated</td>
<td>75.09</td>
<td>74.25</td>
</tr>
</tbody>
</table>
Final Model

- Introduce parameters $\zeta$ and $\tau$ to control the importance of the context factors and of the entity-entity interaction factors.

- **PBoH** model (**Probabilistic Bag of Hyperlinks**):

\[
\log P(e|m, c) \propto \sum_{i=1}^{n} \log P(e_i|m_i) + \zeta \sum_{i=1}^{n} \sum_{w_j \in c_i} \log P(w_j|e_i) + \frac{2\tau}{n-1} \sum_{i<j} \log \left( \frac{P(e_i, e_j)}{P(e_i) P(e_j)} \right)
\]

- mention - entity compatibility
- context - entity interactions
- entity - entity coherence
## Incremental accuracy

<table>
<thead>
<tr>
<th>Baselines</th>
<th>CoNLL test A</th>
<th>CoNLL test B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@MI</td>
<td>R@MA</td>
</tr>
<tr>
<td>Mention-Entity</td>
<td>69.73</td>
<td>69.30</td>
</tr>
<tr>
<td>Uncalibrated</td>
<td>69.77</td>
<td>69.95</td>
</tr>
<tr>
<td>Calibrated</td>
<td>75.09</td>
<td>74.25</td>
</tr>
<tr>
<td>Local Context</td>
<td>82.50</td>
<td>81.56</td>
</tr>
<tr>
<td><strong>PBoH</strong></td>
<td><strong>85.53</strong></td>
<td><strong>85.09</strong></td>
</tr>
</tbody>
</table>

**Table 1:** Accuracy gains of individual *PBoH* components.
MAP inference: \[ e^* = \arg \max_{e \in \mathcal{E}^n} P(e|m, c) \]

- Exact inference - intractable; resort to approximate inference: loopy belief propagation (similar to [Ferragina et al. 2010])

- Fast empirical convergence (typically < 3 iterations, \( \sim 400 \text{ms/doc} \))

- Learning all pairwise parameters \( \rho, \lambda \) was not successful

- Entity candidate pruning:
  - First, top 64 entities based on \( P(e|m) \)
  - Then, keep only top 10 based on \( P(e|m, c) \)
<table>
<thead>
<tr>
<th></th>
<th>ACE2004</th>
<th>CoNLL-Comp.</th>
<th>CoNLL-Test A</th>
<th>CoNLL-Test B</th>
<th>CoNLL-Train</th>
<th>AQUAINT</th>
<th>DBpediaSpotl.</th>
<th>IITB</th>
<th>KORE50</th>
<th>μposts14-Test</th>
<th>μposts14-Train</th>
<th>MSNBC</th>
<th>N3-Reuters-128</th>
<th>N3-RSS-500</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F1@MI</strong></td>
<td>65.83</td>
<td>60.27</td>
<td>59.06</td>
<td>58.32</td>
<td>61.05</td>
<td>60.10</td>
<td>36.61</td>
<td>41.23</td>
<td>34.16</td>
<td>42.43</td>
<td>50.39</td>
<td>75.42</td>
<td>67.95</td>
<td>59.88</td>
</tr>
<tr>
<td><strong>F1@MA</strong></td>
<td>77.63</td>
<td>56.97</td>
<td>53.36</td>
<td>58.03</td>
<td>57.53</td>
<td>58.62</td>
<td>33.25</td>
<td>43.38</td>
<td>30.20</td>
<td>61.08</td>
<td>62.87</td>
<td>73.82</td>
<td>75.52</td>
<td>70.80</td>
</tr>
<tr>
<td>Babelfy</td>
<td>63.20</td>
<td>78.00</td>
<td>75.77</td>
<td>80.36</td>
<td>78.01</td>
<td>72.27</td>
<td>51.05</td>
<td>57.13</td>
<td>73.12</td>
<td>47.20</td>
<td>50.60</td>
<td>78.17</td>
<td>58.61</td>
<td>69.17</td>
</tr>
<tr>
<td></td>
<td>76.71</td>
<td>73.81</td>
<td>71.26</td>
<td>74.52</td>
<td>74.22</td>
<td>73.23</td>
<td>51.97</td>
<td>55.36</td>
<td>69.77</td>
<td>62.11</td>
<td>61.02</td>
<td>75.73</td>
<td>59.87</td>
<td>76.00</td>
</tr>
<tr>
<td>DBpediaSpotlight</td>
<td>70.38</td>
<td>58.84</td>
<td>54.90</td>
<td>57.69</td>
<td>60.04</td>
<td>74.03</td>
<td>69.27</td>
<td>65.44</td>
<td>37.59</td>
<td>56.43</td>
<td>56.26</td>
<td>69.27</td>
<td>56.44</td>
<td>57.63</td>
</tr>
<tr>
<td></td>
<td>80.02</td>
<td>60.59</td>
<td>54.11</td>
<td>61.34</td>
<td>62.23</td>
<td>73.13</td>
<td>67.23</td>
<td>62.81</td>
<td>32.90</td>
<td>71.63</td>
<td>67.99</td>
<td>69.82</td>
<td>58.77</td>
<td>65.03</td>
</tr>
<tr>
<td>Dexter</td>
<td>18.72</td>
<td>48.46</td>
<td>45.44</td>
<td>48.59</td>
<td>49.25</td>
<td>38.28</td>
<td>26.70</td>
<td>28.53</td>
<td>17.20</td>
<td>31.27</td>
<td>35.21</td>
<td>36.86</td>
<td>32.74</td>
<td>31.11</td>
</tr>
<tr>
<td></td>
<td>16.97</td>
<td>45.29</td>
<td>42.17</td>
<td>46.20</td>
<td>45.85</td>
<td>38.15</td>
<td>22.75</td>
<td>28.48</td>
<td>12.54</td>
<td>44.02</td>
<td>39.42</td>
<td>31.85</td>
<td>33.55</td>
<td>32.44</td>
</tr>
<tr>
<td>EntityClassifier.eu</td>
<td>12.74</td>
<td>46.6</td>
<td>44.13</td>
<td>44.02</td>
<td>47.83</td>
<td>21.67</td>
<td>22.59</td>
<td>18.46</td>
<td>27.97</td>
<td>29.12</td>
<td>32.69</td>
<td>41.24</td>
<td>28.4</td>
<td>21.77</td>
</tr>
<tr>
<td></td>
<td>12.3</td>
<td>42.86</td>
<td>42.36</td>
<td>41.31</td>
<td>43.36</td>
<td>19.59</td>
<td>18.0</td>
<td>19.54</td>
<td>25.2</td>
<td>39.53</td>
<td>38.41</td>
<td>40.3</td>
<td>24.84</td>
<td>22.2</td>
</tr>
<tr>
<td>Kea</td>
<td>80.08</td>
<td>73.39</td>
<td>70.9</td>
<td>72.64</td>
<td>74.22</td>
<td>81.84</td>
<td>73.63</td>
<td>72.03</td>
<td>57.95</td>
<td>64.67</td>
<td>63.4</td>
<td>64.67</td>
<td>85.49</td>
<td>63.2</td>
</tr>
<tr>
<td></td>
<td>87.57</td>
<td>73.26</td>
<td>67.91</td>
<td>73.31</td>
<td>74.47</td>
<td>81.27</td>
<td>76.60</td>
<td>70.52</td>
<td>53.17</td>
<td>74.32</td>
<td>87.4</td>
<td>64.45</td>
<td>75.93</td>
<td>70.95</td>
</tr>
<tr>
<td>NERD-ML</td>
<td>54.89</td>
<td>54.62</td>
<td>52.84</td>
<td>52.59</td>
<td>55.55</td>
<td>49.68</td>
<td>46.8</td>
<td>51.08</td>
<td>29.96</td>
<td>38.65</td>
<td>39.83</td>
<td>64.03</td>
<td>54.96</td>
<td>61.22</td>
</tr>
<tr>
<td></td>
<td>72.22</td>
<td>52.35</td>
<td>49.6</td>
<td>51.34</td>
<td>53.23</td>
<td>46.06</td>
<td>45.59</td>
<td>49.91</td>
<td>24.75</td>
<td>57.91</td>
<td>53.74</td>
<td>62.82</td>
<td>62.9</td>
<td>67.3</td>
</tr>
<tr>
<td>TagMe 2</td>
<td>81.93</td>
<td>72.07</td>
<td>69.07</td>
<td>70.62</td>
<td>73.2</td>
<td>76.27</td>
<td>63.31</td>
<td>57.23</td>
<td>57.34</td>
<td>56.81</td>
<td>59.14</td>
<td>75.96</td>
<td>59.32</td>
<td>78.05</td>
</tr>
<tr>
<td></td>
<td>89.09</td>
<td>71.19</td>
<td>66.5</td>
<td>70.38</td>
<td>72.45</td>
<td>75.12</td>
<td>65.1</td>
<td>55.8</td>
<td>54.67</td>
<td>71.66</td>
<td>70.45</td>
<td>77.05</td>
<td>67.55</td>
<td>83.2</td>
</tr>
<tr>
<td>WAT</td>
<td>80.0</td>
<td>83.82</td>
<td>81.82</td>
<td>84.34</td>
<td>84.21</td>
<td>76.82</td>
<td>65.18</td>
<td>61.14</td>
<td>58.99</td>
<td>59.56</td>
<td>61.96</td>
<td>77.72</td>
<td>64.38</td>
<td>68.21</td>
</tr>
<tr>
<td></td>
<td>86.49</td>
<td>83.59</td>
<td>80.25</td>
<td>84.12</td>
<td>84.22</td>
<td>77.64</td>
<td>68.24</td>
<td>59.36</td>
<td>53.13</td>
<td>73.89</td>
<td>72.65</td>
<td>79.08</td>
<td>65.81</td>
<td>76.0</td>
</tr>
<tr>
<td>Wikipedia Miner</td>
<td>77.14</td>
<td>64.72</td>
<td>61.65</td>
<td>60.71</td>
<td>66.48</td>
<td>75.96</td>
<td>62.57</td>
<td>58.59</td>
<td>41.63</td>
<td>54.88</td>
<td>55.93</td>
<td>64.25</td>
<td>60.05</td>
<td>64.54</td>
</tr>
<tr>
<td></td>
<td>86.36</td>
<td>66.17</td>
<td>61.67</td>
<td>63.19</td>
<td>67.93</td>
<td>74.63</td>
<td>61.43</td>
<td>56.98</td>
<td>35.0</td>
<td>69.29</td>
<td>67.0</td>
<td>64.68</td>
<td>66.51</td>
<td>72.23</td>
</tr>
<tr>
<td><strong>PBoH</strong></td>
<td>87.19</td>
<td>86.72</td>
<td>85.63</td>
<td>87.39</td>
<td>86.59</td>
<td>86.64</td>
<td>79.48</td>
<td>62.47</td>
<td>61.70</td>
<td>74.19</td>
<td>73.08</td>
<td>89.54</td>
<td>76.54</td>
<td>71.24</td>
</tr>
<tr>
<td></td>
<td>90.40</td>
<td>86.85</td>
<td>85.48</td>
<td>86.32</td>
<td>87.30</td>
<td>86.14</td>
<td>80.13</td>
<td>55.83</td>
<td>84.48</td>
<td>81.25</td>
<td>89.62</td>
<td>83.31</td>
<td>78.33</td>
<td>71.24</td>
</tr>
</tbody>
</table>

Table 2: Micro and macro F1 scores on the Gerbil\(^1\) (v1.1.4) platform. We highlight the best system and the 2\(^{\text{nd}}\) best system.

\(^1\) [Usbeck et al. 2015]
Conclusions

- Presented a new state of the art entity disambiguation system
- Light-weight probabilistic model based on simple data statistics - scalable to massive amounts of data.
- Plug-in parameter estimators
- Loopy belief propagation inference technique
- Very good generalization performance across many datasets

Future work:
- Alleviate data sparseness using low-dimensional entity vector representations
- Joint MD and ED

Code soon: github.com/dalab/pboh-entity-linking
References

Ferragina, Paolo and Scaiella, Ugo (2010)
Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)

Usbeck, Ricardo et al. (2015)
GERBIL: General Entity Annotator Benchmarking Framework

Blanco, Roi et al. (2015)
Fast and Space-Efficient Entity Linking in Queries

Hoffart, Johannes et al. (2011)
Robust disambiguation of named entities in text

Han, Xianpei et al. (2011)
Collective entity linking in web text: a graph-based method

Guo, Zhaochen and Barbosa, Denilson (2014)
Robust Entity Linking via Random Walks

Milne, David and Witten, Ian H (2008)
Learning to link with wikipedia

Ratinov, Lev et al. (2011)
Local and global algorithms for disambiguation to wikipedia
Thank you!