Improved Decipherment of Homophonic Ciphers

Malte Nuhn and **Julian Schamper** and **Hermann Ney**
Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University, Aachen, Germany
<surname>@cs.rwth-aachen.de

Abstract

In this paper, we present two improvements to the beam search approach for solving homophonic substitution ciphers presented in Nuhn et al. (2013): An improved rest cost estimation together with an optimized strategy for obtaining the order in which the symbols of the cipher are deciphered reduces the beam size needed to successfully decipher the Zodiac-408 cipher from several million down to less than one hundred: The search effort is reduced from several hours of computation time to just a few seconds on a single CPU. These improvements allow us to successfully decipher the second part of the famous Beale cipher (see (Ward et al., 1885) and e.g. (King, 1993)): Having 182 different cipher symbols while having a length of just 762 symbols, the decipherment is way more challenging than the decipherment of the previously deciphered Zodiac-408 cipher (length 408, 54 different symbols). To the best of our knowledge, this cipher has not been deciphered automatically before.

1 Introduction

State-of-the-art statistical machine translation systems use large amounts of parallel data to estimate translation models. However, parallel corpora are expensive and not available for every domain.

Decipherment uses only monolingual data to train a translation model: Improving the core decipherment algorithms is an important step for making decipherment techniques useful for training practical machine translation systems.

In this paper we present improvements to the beam search algorithm for deciphering homophonic substitution ciphers as presented in Nuhn et al. (2013). We show significant improvements in computation time on the Zodiac-408 cipher and show the first decipherment of part two of the Beale ciphers.

2 Related Work

Regarding the decipherment of 1:1 substitution ciphers, various works have been published: Most older papers do not use a statistical approach and instead define some heuristic measures for scoring candidate decipherments. Approaches like Hart (1994) and Olson (2007) use a dictionary to check if a decipherment is useful. Clark (1998) defines other suitability measures based on n-gram counts and presents a variety of optimization techniques like simulated annealing, genetic algorithms and tabu search. On the other hand, statistical approaches for 1:1 substitution ciphers are published in the natural language processing community: Ravi and Knight (2008) solve 1:1 substitution ciphers optimally by formulating the decipherment problem as an integer linear program (ILP) while Corlett and Penn (2010) solve the problem using A^* search. Ravi and Knight (2011) report the first automatic decipherment of the Zodiac-408 cipher. They use a combination of a 3-gram language model and a word dictionary. As stated in the previous section, this work can be seen as an extension of Nuhn et al. (2013). We will therefore make heavy use of their definitions and approaches, which we will summarize in Section 3.

3 General Framework

In this Section we recap the beam search framework introduced in Nuhn et al. (2013).

3.1 Notation

We denote the ciphertext with $f_1^N = f_1 \dots f_j \dots f_N$ which consists of cipher

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1764–1768,
October 25-29, 2014, Doha, Qatar.

tokens $f_j \in V_f$. We denote the plaintext with $e_1^N = e_1 \dots e_i \dots e_N$ (and its vocabulary V_e respectively). We define $e_0 = f_0 = e_{N+1} = f_{N+1} = \$$ with "$" being a special sentence boundary token. Homophonic substitutions are formalized with a general function $\phi : V_f \rightarrow V_e$. Following (Corlett and Penn, 2010), cipher functions ϕ, for which not all $\phi(f)$'s are fixed, are called partial cipher functions. Further, ϕ' is said to extend ϕ, if for all $f \in V_f$ that are fixed in ϕ, it holds that f is also fixed in ϕ' with $\phi'(f) = \phi(f)$. The cardinality of ϕ counts the number of fixed f's in ϕ. When talking about partial cipher functions we use the notation for relations, in which $\phi \subseteq V_f \times V_e$.

3.2 Beam Search

The main idea of (Nuhn et al., 2013) is to structure all partial ϕ's into a search tree: If a cipher contains N unique symbols, then the search tree is of height N. At each level a decision about the n-th symbol is made. The leaves of the tree form full hypotheses. Instead of traversing the whole search tree, beam search descents the tree top to bottom and only keeps the most promising candidates at each level. Practically, this is done by keeping track of all partial hypotheses in two arrays H_s and H_t. During search all allowed extensions of the partial hypotheses in H_s are generated, scored and put into H_t. Here, the function EXT_ORDER (see Section 5) chooses which cipher symbol is used next for extension, EXT_LIMITS decides which extensions are allowed, and SCORE (see Section 4) scores the new partial hypotheses. PRUNE then selects a subset of these hypotheses. Afterwards the array H_t is copied to H_s and the search process continues with the updated array H_s. Figure 1 shows the general algorithm.

```
1: function BEAM_SEARCH(EXT_ORDER)
2:     init sets H_s, H_t
3:     CARDINALITY = 0
4:     H_s.ADD((∅, 0))
5:     while CARDINALITY < |V_f| do
6:         f = EXT_ORDER[CARDINALITY]
7:         for all φ ∈ H_s do
8:             for all e ∈ V_e do
9:                 φ' := φ ∪ {(e, f)}
10:                if EXT_LIMITS(φ') then
11:                    H_t.ADD(φ',SCORE (φ'))
12:                end if
13:            end for
14:        end for
15:        PRUNE(H_t)
16:        CARDINALITY = CARDINALITY + 1
17:        H_s = H_t
18:        H_t.CLEAR()
19:    end while
20:    return best scoring cipher function in H_s
21: end function
```

Figure 1: The general structure of the beam search algorithm for decipherment of substitution ciphers as presented in Nuhn et al. (2013). This paper improves the functions SCORE and EXT_ORDER.

4 Score Estimation

The score estimation function is crucial to the search procedure: It predicts how good or bad a partial cipher function ϕ might become, and therefore, whether it's worth to keep it or not.

To illustrate how we can calculate these scores, we will use the following example with vocabularies $V_f = \{A, B, C, D\}$, $V_e = \{a, b, c, d\}$, extension order (B, C, A, D), and cipher text[1]

$$f_1^N = \texttt{\$ ABDD CABC DADC ABDC \$}$$

[1]We include blanks only for clarity reasons.

and partial hypothesis $\phi = \{(A, a), (B, b)\}$. This yields the following partial decipherment

$$\phi(f_1^N) = \texttt{\$ ab.. .ab. .a.. ab.. \$}$$

The score estimation function can only use this partial decipherment to calculate the hypothesis' score, since there are not yet any decisions made about the other positions.

4.1 Baseline

Nuhn et al. (2013) present a very simple rest cost estimator, which calculates the hypothesis' score based only on fully deciphered n-grams, i.e. those parts of the partial decipherment that form a contiguous chunk of n deciphered symbols. For all other n-grams containing not yet deciphered symbols, a trivial estimate of probability 1 is assumed, making it an admissible heuristic. For the above example, this baseline yields the probability $p(a|\$) \cdot p(b|a) \cdot 1^4 \cdot p(b|a) \cdot 1^6 \cdot p(b|a) \cdot 1^2$. The more symbols are fixed, the more contiguous n-grams become available. While being easy and efficient to compute, it can be seen that for example the single "a" is not involved in the computation of

the score at all. In practical decipherment, like e.g. the Zodiac-408 cipher, this forms a real problem: While making the first decisions—i.e. traversing the first levels of the search tree—only very few terms actually contribute to the score estimation, and thus only give a very coarse score. This makes the beam search "blind" when not many symbols are deciphered yet. This is the reason, why Nuhn et al. (2013) need a large beam size of several million hypotheses in order to not lose the right hypothesis during the first steps of the search.

4.2 Improved Rest Cost Estimation

The rest cost estimator we present in this paper solves the problem mentioned in the previous section by also including lower order n-grams: In the example mentioned before, we would also include unigram scores into the rest cost estimate, yielding a score of $p(a|\$) \cdot p(b|a) \cdot 1^3 \cdot p(a) \cdot p(b|a) \cdot 1^2 \cdot p(a) 1^2 \cdot p(a) \cdot p(b|a) \cdot 1^2$. Note that this is not a simple linear interpolation of different n-gram trivial scores: Each symbol is scored only using the maximum amount of context available. This heuristic is non-admissible, since an increased amount of context can always lower the probabilty of some symbols. However, experiments show that this score estimation function works great.

5 Extension Order

Besides having a generally good scoring function, also the order in which decisions about the cipher symbols are made is important for obtaining reliable cost estimates. Generally speaking we want an extension order that produces partial decipherments that contain useful information to decide whether a hypothesis is worth being kept or not as early as possible.

It is also clear that the choice of a good extension order is dependent on the score estimation function SCORE. After presenting the previous state of the art, we introduce a new extension order optimized to work together with our previously introduced rest cost estimator.

5.1 Baseline

In (Nuhn et al., 2013), two strategies are presented: One which at each step chooses the most frequent remaining cipher symbol, and another, which greedily chooses the next symbol to maximize the number of contiguously fixed n-grams in the ciphertext.

LM order	Perplexity	
	Zodiac-408	Beale Pt. 2
1	19.49	18.35
2	14.09	13.96
3	12.62	11.81
4	11.38	10.76
5	11.19	9.33
6	10.13	8.49
7	10.15	8.27
8	9.98	8.27

Table 1: Perplexities of the correct decipherment of Zodiac-408 and part two of the Beale ciphers using the character based language model used in beam search. The language model was trained on the English Gigaword corpus.

5.2 Improved Extension Order

Each partial mapping ϕ defines a partial decipherment. We want to choose an extension order such that *all possible* partial decipherments following this extension order are as informative as possible: Due to that, we can only use information about *which* symbols will be deciphered, not their actual decipherment. Since our heuristic is based on n-grams of different orders, it seems natural to evaluate an extension order by counting how many contiguously deciphered n-grams are available: Our new strategy tries to find an extension order optimizing the weighted sum of contiguously deciphered n-gram counts[2]

$$\sum_{n=1}^{N} w_n \cdot \#_n.$$

Here n is the n-gram order, w_n the weight for order n, and $\#_n$ the number of positions whose maximum context is of size n.

We perform a beam search over all possible enumerations of the cipher vocabulary: We start with fixing only the first symbol to decipher. We then continue with the second symbol and evaluate all resulting extension orders of length 2. In our experiments, we prune these candidates to the 100 best ones and continue with length 3, and so on.

Suitable values for the weights w_n have to be chosen. We try different weights for the different

[2]If two partial extension orders have the same score after fixing n symbols, we fall back to comparing the scores of the partial extension orders after fixing only the first $n-1$ symbols.

i_{02}	h_{08}	a_{03}	v_{01}	e_{05}	d_{09}	e_{07}	p_{03}	o_{07}	s_{10}	i_{11}	t_{03}	e_{14}	d_{03}	i_{03}	n_{05}	t_{06}	h_{01}	e_{13}	c_{04}	o_{10}	u_{01}	n_{01}	t_{04}	y_{01}
o_{12}	f_{04}	b_{04}	e_{15}	d_{09}	f_{03}	o_{04}	r_{06}	d_{04}	a_{07}	b_{07}	o_{09}	u_{03}	t_{13}	f_{01}	o_{01}	u_{08}	r_{05}	m_{03}	i_{08}	l_{09}	e_{14}	s_{06}	f_{01}	r_{05}
o_{07}	m_{04}	b_{06}	u_{02}	f_{04}	o_{10}	r_{07}	d_{01}	s_{11}	i_{03}	n_{02}	a_{06}	n_{03}	e_{05}	x_{01}	c_{03}	a_{01}	v_{01}	a_{03}	t_{10}	i_{13}	o_{03}	n_{05}	o_{08}	r_{06}
v_{01}	a_{08}	u_{03}	l_{01}	t_{11}	s_{12}	i_{04}	x_{01}	f_{01}	e_{01}	e_{03}	t_{02}	b_{06}	e_{07}	l_{02}	o_{11}	w_{06}	t_{08}	h_{08}	e_{15}	s_{06}	u_{04}	r_{06}	f_{04}	a_{10}
											$\vdots$												
p_{04}	a_{14}	p_{01}	e_{07}	r_{05}	n_{02}	u_{02}	m_{02}	b_{01}	e_{14}	r_{05}	o_{03}	n_{05}	e_{15}	d_{10}	e_{01}	s_{01}	c_{01}	r_{01}	i_{03}	b_{05}	e_{06}	s_{08}	t_{01}	h_{08}
c_{04}	e_{10}	x_{01}	a_{14}	c_{07}	t_{02}	l_{09}	o_{12}	c_{02}	a_{04}	l_{09}	i_{13}	t_{02}	y_{01}	o_{02}	f_{03}	t_{07}	h_{02}	e_{11}	v_{01}	a_{10}	r_{07}	l_{07}	t_{11}	s_{09}
o_{04}	t_{01}	h_{03}	a_{06}	t_{04}	n_{03}	o_{06}	d_{05}	i_{13}	f_{02}	f_{03}	i_{03}	c_{04}	u_{07}	l_{09}	t_{02}	y_{01}	w_{04}	i_{12}	l_{01}	l_{02}	b_{03}	e_{01}	h_{02}	a_{09}
d_{10}	i_{07}	n_{06}	f_{01}	i_{13}	n_{01}	d_{10}	i_{03}	n_{05}	g_{04}	i_{03}	t_{05}												

Table 2: Beginning and end of part two of the Beale cipher. Here we show a relabeled version of the cipher, which encodes knowledge of the gold decipherment to assign reasonable names to all homophones. The original cipher just consists of numbers.

orders on the Zodiac-408 cipher with just a beam size of 26. With such a small beam size, the extension order plays a crucial role for a successful decipherment: Depending on the choice of the different weights w_n we can observe decipherment runs with 3 out of 54 correct mappings, up to 52 out of 54 mappings correct. Even though the choice of weights is somewhat arbitrary, we can see that generally giving higher weights to higher n-gram orders yields better results.

We use the weights $w_1^8 = (0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 3.0)$ for the following experiments. It is interesting to compare these weights to the perplexities of the correct decipherment measured using different n-gram orders (Table 5). However, at this point we do not see any obvious connection between perplexities and weights w_n, and leave this as a further research direction.

6 Experimental Evaluation

6.1 Zodiac Cipher

Using our new algorithm we are able to decipher the Zodiac-408 with just a beam size of 26 and a language model order of size 8. By keeping track of the gold hypothesis while performing the beam search, we can see that the gold decipherment indeed always remains within the top 26 scoring hypotheses. Our new algorithm is able to decipher the Zodiac-408 cipher in less than 10s on a single CPU, as compared to 48h of CPU time using the previously published heuristic, which required a beam size of several million. Solving a cipher with such a small beam size can be seen as "reading off the solution".

6.2 Beale Cipher

We apply our algorithm to the second part of the Beale ciphers with a 8-gram language model.

Compared to the Zodiac-408, which has length 408 while having 54 different symbols (7.55 observations per symbol), part two of the Beale ciphers has length 762 while having 182 different symbols (4.18 observations per symbol). Compared to the Zodiac-408, this is both, in terms of redundancy, as well as in size of search space, a way more difficult cipher to break.

Here we run our algorithm with a beam size of 10M and achieve a decipherment accuracy of 157 out of 185 symbols correct yielding a symbol error rate of less than 5.4%. The gold decipherment is pruned out of the beam after 35 symbols have been fixed.

We also ran our algorithm on the other parts of the Beale ciphers: The first part has a length 520 and contains 299 different cipher symbols (1.74 observations per symbol), while part three has length 618 and has 264 symbols which is 2.34 observations per mapping. However, our algorithm does not yield any reasonable decipherments. Since length and number of symbols indicate that deciphering these ciphers is again more difficult than for part two, it is not clear whether the other parts are not a homophonic substitution cipher at all, or whether our algorithm is still not good enough to find the correct decipherment.

7 Conclusion

We presented two extensions to the beam search method presented in (Nuhn et al., 2012), that reduce the search effort to decipher the Zodiac-408 enormously. These improvements allow us to automatically decipher part two of the Beale ciphers. To the best of our knowledge, this has not been

done before. This algorithm might prove useful when applied to word substitution ciphers and to learning translations from monolingual data.

Acknowledgements

The authors thank Mark Kozek from the Department of Mathematics at Whittier College for challenging us with a homophonic cipher he created. Working on his cipher led to developing the methods presented in this paper.

References

Andrew J. Clark. 1998. *Optimisation heuristics for cryptology*. Ph.D. thesis, Faculty of Information Technology, Queensland University of Technology.

Eric Corlett and Gerald Penn. 2010. An exact A* method for deciphering letter-substitution ciphers. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1040–1047, Uppsala, Sweden, July. The Association for Computer Linguistics.

George W. Hart. 1994. To decode short cryptograms. *Communications of the Association for Computing Machinery (CACM)*, 37(9):102–108, September.

John C. King. 1993. A reconstruction of the key to beale cipher number two. *Cryptologia*, 17(3):305–317.

Malte Nuhn, Arne Mauser, and Hermann Ney. 2012. Deciphering foreign language by combining language models and context vectors. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 156–164, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Malte Nuhn, Julian Schamper, and Hermann Ney. 2013. Beam search for solving substitution ciphers. In *Annual Meeting of the Assoc. for Computational Linguistics*, pages 1569–1576, Sofia, Bulgaria, August.

Edwin Olson. 2007. Robust dictionary attack of short simple substitution ciphers. *Cryptologia*, 31(4):332–342, October.

Sujith Ravi and Kevin Knight. 2008. Attacking decipherment problems optimally with low-order n-gram models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 812–819, Honolulu, Hawaii. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011. Bayesian inference for Zodiac and other homophonic ciphers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 239–247, Portland, Oregon, June. Association for Computational Linguistics.

James B Ward, Thomas Jefferson Beale, and Robert Morriss. 1885. *The Beale Papers*.