Communicative Grounding of Analogical Explanations in Dialogue: A Corpus Study of Conversational Management Acts and Statistical Sequence Models for Tutoring through Analogy

Jorge Del-Bosque-Trevino[1], Julian Hough[2], Matthew Purver[3]
Computational Linguistics Lab
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
[1]Education Lab [2]Human Interaction Lab [3]Jožef Stefan Institute
{j.delbosque, j.hough, m.purver}@qmul.ac.uk

Abstract

We present a conversational management act (CMA) annotation schema for one-to-one tutorial dialogue sessions where a tutor uses an analogy to teach a student a concept. CMAs are more fine-grained sub-utterance acts compared to traditional dialogue act mark-up. The schema achieves an inter-annotator agreement (IAA) Cohen Kappa score of at least 0.66 across all 10 classes. We annotate a corpus of analogical episodes with the schema and develop statistical sequence models from the corpus which predict tutor content related decisions, in terms of the selection of the analogical component (AC) and tutor conversational management act (TCMA) to deploy at the current utterance, given the student's behaviour. CRF sequence classifiers perform well on AC selection and robustly on TCMA selection, achieving respective accuracies of 61.9% and 56.3% on a cross-validation experiment over the corpus.

1 Introduction

The motivation for our work is two-fold; firstly it derives from our interest to investigate which conversational management acts (CMAs) are prevalent in tutorial explanations which use analogies as a pedagogical strategy, and secondly, our interest to determine what the optimal computational model is for selecting the appropriate analogical components (ACs) and tutor conversational management acts (TCMAs) for an artificial tutoring agent. These findings represent a milestone towards our end goal of building a Conversational AI system which is capable of tutoring in introductory computer science topics on a speech based modality.

The paper is presented in the following sections: in Section 2 we present theoretical and empirical foundations on dialogue acts (DAs), tutorial dialogue, hidden markov models (HMMs) and conditional random fields (CRF) modelling; Section 3 presents the first contribution of this paper, which is the development of an annotation schema of tutor conversational management acts (TCMAs) and student conversational management acts (SCMAs); Section 4 presents a corpus study on dialogues annotated using the developed schema; Sections 5 and 6 present the second contribution of this investigation, (i) predictive models on analogical component (AC) uttered by the tutor and (ii) models predicting TCMAs, followed by concluding the findings and future work in Section 7.

2 Background

2.1 Tutorial Dialogue Modelling

The problem of determining interaction patterns which increase effectiveness and maximise learning gains in tutorial sessions is a significant area of interest within the field of face-to-face (Porayska-Pomsta and Mellish, 2013; Rus et al., 2017; VanLehn et al., 2003), computer-mediated tutoring (Siler and VanLehn, 2009) and intelligent tutoring systems (ITS) (Aleven and Koedinger, 2002; Alizadeh et al., 2015; Di Eugenio et al., 2013). Previous methods grounded in tutoring and natural language theories have shown how Hidden Markov Models can be learned from human-human corpora and have been applied to discover tutorial strategies (Boyer et al., 2009, 2011). However, such sequence modelling methods have not been applied to the pedagogical strategy of analogical explanations.

2.2 Dialogue Act Theory, Taxonomies and Annotation Schemata

As part of the development process of our annotation schema, we integrate speech act theory

Proceedings of the Conference on Reasoning and Interaction (ReInAct), pages 23–31

(Searle, 1965), discourse markers theory (Schiffrin, 1987; Fraser, 1999; Schourup, 1999), grounding in communication theory (Clark and Brennan, 1991), the DAMSL annotation schema (Allen and Core, 1997) and 3 annotation schemata developed to test pedagogical theories in tutorial dialogues (Di Eugenio et al., 2009, 2013; Alizadeh et al., 2015). We use a previously annotated corpus on computer science tutorial dialogues (CSTD) (Di Eugenio et al., 2009) . The CSTD corpus has been coded with various annotation schemata which were specifically developed to test hypotheses in the pedagogical domain (Di Eugenio et al., 2009, 2013; Alizadeh et al., 2015). We take these previous schemata as a point of departure and develop our own schema explicitly created for investigating grounding analogies in tutorial dialogues. This new fine-grained schema, allows us to further research the phenomenon of AC and TCMA deployment at a higher resolution, as explained in sections 3 and 4.

2.3 Analogical Explanations

Analogical explanations contain *base* and *target* components (Gentner, 1983) and CMAs, as shown in a sample of an analogical explanation from the Computer Science Tutorial Dialogue (CSTD) Corpus (Di Eugenio et al., 2009) in Table 1. The *base* is formatted in **bold** type and the *target* in **underlined bold** type. These ACs are communicated by the tutor using TCMAs, which are *underlined and italicised* in the example. The student participant in the dialogue issues utterances containing student conversational management acts (SCMA), which are *italicised*. Del-Bosque-Trevino et al. (2020) showed that ACs occur in tutorial dialogues in regular patterns which resemble the *semantic wave* (Maton, 2013; Curzon et al., 2018) where the tutor begins their explanations with high semantic density in the *target* domain, descends to lower density in the *base* domain and returns to the *target* domain. The challenge of this investigation is to develop predictive models that would enable an artificial tutoring agent to select the optimal AC and TCMA while interacting with a student.

3 Annotation Schema

To develop a schema of Tutor Conversational Management Acts (TCMA) and Student Conversational Management Acts (SCMA), we selected 600 utterances from analogical episodes

Tutor	*so*, the way we can think of a **stack** is, kind of like a **brick wall**, *right?*
Tutor	there is *uh*, **we lay a brick down, and every time we put something else on the stack we always lay on top of the previous one.**
Tutor	*right?*
Student	*yeah.*
Tutor	**when we build a wall we build the bottom up.**
Student	*yeah, thats true.*

Table 1: Tutor and Student Dialogue depicting analogical components (ACs) and Conversational Management Acts (CMAs). **Underlined bold** for Target components, **bold** for Base components), *underlined italics* for Tutor Conversational Management Acts (TCMAs) and *italics* for Student Conversational Management Acts (SCMAs)

.

from the CSTD Corpus (Di Eugenio et al., 2009) of tutorial dialogues. We achieve Cohen's Kappa Inter-Annotator Agreement (IAA) of at least 0.66 for every class as shown in Table 4. An overview of TCMA and SCMA classes annotated with a description and example for each one can be seen in Tables 2 and 3.

3.1 Annotators and Tags

The first author and a second annotator participated in the annotation exercise to verify the schema. Regarding TCMAs, we split what was previously coded as PT (Tutor Prompt) into five TCMAs; Q (Question Response), ER (Eliciting Response), FIM (Floor Initiating or Maintaining), ABK (Acknowledge Backchannel, DB (Diagnosing Base). Their descriptions and examples are shown in Table 2. Concerning SCMAs, we split what was originally annotated from previous experiments as SI (Student Initiative) into five SCMAs; BK (Backchannel), BKR (Backchannel through Repetition), SCC (Student Collaborative Completion, E (Exclamation), Q (Question). The description of the Student's tags and examples are presented in Table 3.

4 Corpus Study

The CSTD Corpus (Di Eugenio et al., 2009) contains 54 sessions of annotated tutorial dialogues. A sub-corpus composed of the 138 sequences of utterances marked as analogical episodes was

Tutor Conversational Management Act	**Description**	**Example**
Question Response (QR)	The tutor responds to a question asked by the student.	Student: "and these two lists are totally separate?" Tutor: "yeah, these are totally separate."
Elicit Response (ER)	Questions by the tutor to test if the student understands what he or she is talking about, or to see if the student is paying attention.	"it's stored +// it stores information, okay?" "right?"
Diagnosing Base (DB)	The tutor asks a question to check if the student is familiar with the base concept of an analogy.	"are you familiar with Legos?"
Floor Initiating or Floor Maintaining Discourse Marker (FIM)	Floor initiating in one utterance. Floor maintaining (the expression is located in the middle of the utterance)	"uh are you familiar with Legos?" "um what the stack is # is a way to hold objects such as uh integers or numbers, letters, words." "okay, a binary search is like a family tree" "now if you notice, there is no. . . " "so now if I'm go again. . . " "so we could push..."
Acknowledge Backchannel (ABK)	Appears immediately after the student issues a "Backchannel"	Student: Right. [BK] Tutor: "Right? [ABK]

Table 2: Tutor Conversational Management Acts (TCMA)

Student Conversational Management Act	**Description**	**Example**
Backchannel (BK)	The student issues a backchannel.	"uh huh", "yeah", "mmhm", "right", "OK","okay"
Backchannel through Repetition (BR)	The student backchannels through repetition.	Tutor: . . . the head is for the next line. Student: line ok.
Student Compound Contribution (SCC)	The student takes up the turn and tries to complete an utterance issued by the tutor. The student pauses his or her utterance an waits for the tutor to complete it.	Tutor: but we need, we can't just say, Bob, point to Greg. Student: "we need to know."
Question (Q)	The student poses a question spontaneously.	Tutor: "Good, remember that formula" Student: "how does the lightbulb's resistance figure in the equations?"
Exclamation (E)	The student makes an exclamation	"oh", "ah"

Table 3: Student Conversational Management Acts (SCMA)

QR	**DB**	**ER**	**FIM**	**ABK**	**Q**	**BK**	**BR**	**SCC**	**E**
0.66	1.00	0.84	0.70	0.81	1.00	0.97	0.79	0.66	1.00

Table 4: Conversational Management Acts IAA

	B	**T**	**BT**	**Row Totals**
SCMA at t-1	300 (46.01%)	201 (30.83%)	151(23.16%)	652 (100%)
No SCMA at t-1	1200 (50.27%)	771 (32.29%)	416 (17.43%)	2387 (100%)
Column Totals	1501 (49.34%)	974 (32.02%)	567 (18.64%)	**3039 (100%)**

Table 5: Student Conversational Management Acts Chi-Square Crosstable

subsequently annotated in (Del-Bosque-Trevino et al., 2020) using an annotation schema grounded on Structure-mapping theory (Gentner, 1983) where each utterance has an annotation of either a *base* component (B), *target* (T) or both (BT). We took that sub-corpus and annotated each utterance with every CMA tag that could apply to it as per the schema described above. A total of 3039 tutor utterances and 777 student utterances were annotated [1].

4.1 Descriptive Statistics

Students and tutors can produce utterances which contain more than one single tag, resulting in compound tags. Table 6 contains the distribution of the student single and compound tags and Table 7 contains the distribution of the tutor single and compound tags.

Tag	Frequency	%
No SCMA	77	9.91%
Single tags		
BK	575	74.00%
Q	47	6.05%
SCC	29	3.73%
E	17	2.19%
BR	5	0.64%
Compound tags		
BK&E	12	1.54%
BK&SCC	5	0.64%
BK&BR	4	0.51%
BK&Q	2	0.26%
SCC&Q	2	0.26%
BR&SCC	1	0.13%
BR&E	1	0.13%
Total	777	100%

Table 6: Frequency Distribution of single and compound CMA tags over all student Utterances.

The dominant class (74%) within SCMAs is BK. Such a high-frequency rate motivated us to analyse the annotated corpus to split it into finer-grained subclasses. After a thorough examination of every instance we found no relevant features that could justify the subdivision of the BK tag. This finding is consistent with the fact that the modelled corpus is tutor initiated / led dialogue.

Over half of the tutor utterances did not contain a TCMA, and the most common tag is a floor-initiating or maintaining discourse marker (FIM), followed by an eliciting response (ER) and acknowledging backchannel (ABK). While rarer, Question Responses (QR), which are important for learning through dialogue, accounted for 1.22%.

[1]The annotations will be available in http://www.delbosque.co/ReInAct2021

4.2 Test of dependence between the presence of SCMA and the AC uttered by the tutor

With the purpose of determining if the production of a SCMA influenced tutor decisions, we conducted a test of independence. Specifically, we tested whether the tutor's selection of analogical component (base, target or both) depended on the presence of a SCMA immediately preceding that utterance. To test this dependence, we performed a chi-square χ^2 test of independence between the type of AC uttered by the tutor at utterance *t* and the presence of an SCMA at utterance *t-1*.

Tag	Frequency	%
No TCMA	1692	55.68%
Single tags		
FIM	707	23.26%
ER	336	11.06%
ABK	137	4.51%
QR	37	1.22%
DB	9	0.30%
Compound tags		
FIM&ER	87	2.86%
ABK&FIM	17	0.56%
FIM&QR	5	0.16%
FIM&ER&ABK	4	0.13%
ABK&ER	4	0.13%
FIM&DB	3	0.10%
ABK&QR	1	0.03%
Total	3039	100%

Table 7: Frequency distribution of single and compound CMA tags over all tutor utterances.

The relation between these variables was significant χ^2 (1, N = 3039) = 11.2474, p = .0036. The result indicates that the tutor decision on whether to produce an utterance with the base, target or both analogical components depends on the presence or absence of a SCMA before the tutor takes or retakes his turn. A cross-table is presented in Table 5 showing a considerable difference between AC of the type BT preceded by a SCMA (23.16%) versus its absence (17.43%). The AC of type B also shows a notable difference changing from 46.01% when it is preceded by a SCMA to 50.27% when it is not. Type T of analogical component presents a variation to a lesser degree compared to the rest of the components, changing from 30.83% to 32.29%. In terms of the tutor's decision, the tutor is more likely to use an AC of type B if there is no student contribution, compared to when there has been one. Conversely, the tutor is more likely to produce an utterance with a Target

component in it (T or BT) if there has been a student contribution as in the previous utterance.

5 Predicting a Tutor's selection of Analogical Component Using Sequence Modelling

Given the eventual goal of building an artificial tutor which can ground analogies with a tutee over time, we use the corpus described above to train two types of sequence model: (i) the tutor's sequential decisions of selecting an appropriate analogical component (AC) and (ii) the tutor selection of the appropriate conversational management act (TCMA) for their next utterance. We test both a generative sequence model, a Hidden Markov Model (HMM) classifier, and a discriminative model, a Conditional Random Field (CRF) classifier. We use Markovian sequence models due to the fact there is a regularity in analogical structure following the semantic wave (Maton, 2013; Curzon et al., 2018). Due to the relatively small size of the possible training data (138 sequences, 3039 tutor utterances), we do not use neural net based sequence architectures in this work. The first type of decision we model is predicting the AC to be communicated in the next tutor utterance from the 3 classes of a Base component (B), Target (T) or both (BT). Our AC prediction models all take as input a sequence of elements which represent the student's last utterance after the most recent tutor utterance (so this could be no utterance ($\varnothing$) if the tutor continues to hold the floor). We are in effect modelling the decision a Dialogue Manager would take in a Dialogue System to select the analogical component to communicate in the next utterance. We set up an HMM and CRF to do this prediction as explained in the following sections.

5.1 HMM for predicting tutor ACs

Our HMM implementation follows that of a first order Maximum Likelihood Estimate (MLE) HMM as described in Chapter 8 of (Jurafsky and Martin, 2020). The model is trained to predict the most likely Analogical Component (AC) at time *t*, taking as an input observation the preceding Student Conversational Management Act (SCMA) type, using a combination of transition and emission probabilities obtained from the corpus briefly described below.

5.1.1 Transition and emission probabilities

We obtain our transition probabilities for the AC tag at the current time step t conditioned on the AC at the previous timestep from our data as a first order (i.e. bigram) MLE markov model $P(ac_t|ac_{t-1})$. We compute the maximum likelihood estimate of this transition probability as the ratio of the counts of the times the first tag occurs in the training data, over how many times the first tag is followed by the second:

$$P(ac_t|ac_{t-1}) = \frac{C(ac_{t-1}, ac_t)}{C(ac_{t-1})}$$

For example, in the whole of our corpus, BT occurs 567 times, of which it is followed by T 112 times, giving an MLE estimate of:

$$P(T|BT) = \frac{C(BT, T)}{C(BT)} = \frac{112}{567} = 0.1975$$

For the emission probabilities we model the likelihood of an observation o_t being generated from a state q_i. In our model, our observations $O = o_1, o_2, o_T$ consist of a sequence of SCMA types, drawn from a vocabulary V of size 13 consisting of the single and compound tags in Table 6 in addition to no observed student utterance ($\varnothing$). The states are the underlying AC component at that timestep.

We compute the MLE emission probability of a given SCMA being observed at time t given an underlying AC state by the ratio of the counts of the occurrences of that AC at t having that SCMA observation in the corpus, as in:

$$P(scma_t|ac_t) = \frac{C(ac_t, scma_t)}{C(ac_t)}$$

For example in our corpus, T occurs 968 times, of which a BK is observed 153 times, for an MLE estimate of:

$$P(BK|T) = \frac{C(T, BK)}{C(T)} = \frac{153}{968} = 0.1580$$

5.1.2 Viterbi decoding

Once the transition and emission probabilities have been obtained through the counts on the corpus, at testing time, these are used in tandem to decode sequences of SCMAs to give the most likely sequence of AC states $ac_{1...t}$. In our HMM, this is formulated in the standard way as in Fig. 1.

$$\arg\max_{ac_{1\ldots t}} P(ac_{1\ldots t}|scma_{1\ldots t}) = \arg\max_{ac_{1\ldots t}} P(ac_{1\ldots t})P(scma_{1\ldots t}|ac_{1\ldots t}) \quad (1)$$

$$= \arg\max_{ac_{1\ldots t}} \prod_{i=1}^{t} P(ac_i|ac_{i-1})P(scma_i|ac_i) \quad (2)$$

Figure 1: Full HMM model for decoding input SCMA sequences to ACs.

To avoid computing all possible state sequences $ac_{1\ldots t}$ we use the Viterbi algorithm to create a state trellis of length t, where the most likely sequence of ACs is computed dynamically.

Fig. 2 shows an intuition of this lattice for the sequence $\varnothing \rightarrow BK \rightarrow \varnothing \rightarrow SCC \rightarrow \varnothing \rightarrow BR \rightarrow E \rightarrow Q$. The illustration also depicts how the semantic wave (Maton, 2013; Curzon et al., 2018) is created while the conversation unfolds over time as the explanation begins with high semantic density, drops in the middle of the episode to the base component of the analogy, then finally rises again.

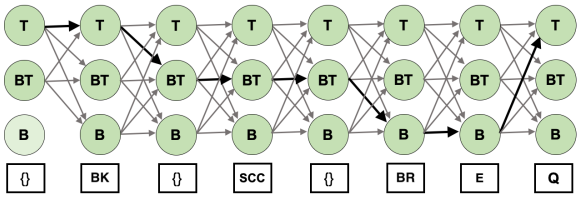

Figure 2: A sketch of the trellis for an SCMA sequence showing the possible tags for each SCMA and highlighting in bold the path corresponding to the correct and most likely tag sequence through the hidden states.

5.1.3 BEIS factoring of AC tags

We also trial a variant of AC states which gives more structure, enriching the tags with BEIS (i.e. Beginning, End, Inside, Single) information, expanding from 3 AC tag types to 12. With the BEIS notation, an example sequence is as follows:

$$S \rightarrow Ts \rightarrow BTb \rightarrow BTi \rightarrow BTi \rightarrow BTi \rightarrow$$

$$BTe \rightarrow Bs \rightarrow BTb \rightarrow BTi \rightarrow BTi \rightarrow$$

$$BTe \rightarrow Ts \rightarrow E$$

(Analogical Episode with BEIS tags)

For evaluation purposes, the decoded BEIS states would then be converted back into the pure AC states in {B, T, BT}.

5.2 CRF Model for Predicting tutor ACs

An alternative discriminative sequence model, a Conditional Random Field (CRF) was also experimented with to predict the most likely AC sequence more directly from the data, i.e. going directly arg $\max_{ac_{1\ldots t}} P(ac_{1\ldots t}|scma_{1\ldots t})$. This was used in order to exploit more fine-grained features beyond simply the SCMA type, to more features of the student utterances. The features used for each time step t relating to the preceding student utterance (if present) were:

- *scma* type
- Lexical features as a Bag-of-Words (BoW) representation of the utterance.
- Utterance length.
- Presence of a question mark in the transcript of the utterance.

Different pre-processing techniques were trialled in removing stop-words and using lemmas rather than words after automatic lemmatization. As per the HMM model, training both using pure AC tags for labels and using the BEIS enriched tags with conversion back to the pure AC tags was trialled.

5.3 Experimental set-up

We set-up an 8-fold cross-validation experiment for all models over the 138 analogical episodes. We evaluate for the accuracy of AC prediction in terms of *accuracy*, *weighted F1 score* and *F1 macro score* (average performance across the three classes).

Both classifiers were implemented in Python. The HMM transition and emission probabilites were obtained on the training part of each fold via NLTK conditional probability distribution objects, and the Viterbi code is adapted from Katrin Erk's HMM Python worksheet.[2] The CRF implementation used was the NLTK CRF Tagger[3] with default settings. The feature extraction

[2] http://www.katrinerk.com/courses/python-worksheets/hidden-markov-models-for-pos-tagging-in-python

[3] https://www.nltk.org/_modules/nltk/tag/crf.html

Model	Tags	Features	F1 macro	F1 weighted	Accuracy
Random	Full states	-	0.3267	0.3474	0.3356
HMM	Full states	SCMA type	0.3985	0.5014	0.5637
HMM	BEIS states	SCMA type	**0.4461**	0.5416	0.5969
CRF	Full states	SCMA type + lexical	0.3968	0.5003	0.5627
CRF	BEIS states	SCMA type + lexical	0.4407	**0.5490**	**0.6186**

Table 8: Performance for Tagging Analogical Component (AC) in a cross-validation.

Model	Tags	Features	B	BT	T
Random	Full states	-	0.3885	0.2526	0.3390
HMM	Full states	SCMA type	0.6837	0.0000	0.5118
HMM	BEIS states	SCMA type	0.7149	0.0817	0.5418
CRF	Full states	SCMA type + lexical	0.6868	0.0000	0.5036
CRF	BEIS states	SCMA type + lexical	0.7300	0.0069	0.5836

Table 9: Breakdown of Performance (F1 weighted average) for AC in a cross-validation

function for each time-step was adjusted to extract the features for each utterance described above.

5.4 Results

The results for tagging analogical components with our HMM and CRF models are presented in Table 8. As shown in the table, The HMM trained on AC substates (BEIS) yields the highest *f-1 macro average score*. Nevertheless the optimal model is the CRF trained with AC substates (BEIS), yielding the highest *f-1 weighted average score* of 0.5490 and *accuracy* of 0.6186. The table also demonstrates how the model is capable of easily beating a random baseline of a *f-1 weighted average score* of 0.3474. The best CRF model is trained using the features of SCMA type, lexical, stopword removal, student utterance length, lemmatisation, and the binary feature of whether the utterance contains a question mark or not. The breakdown results of predicting either B, BT or T components, as shown in Table 9 highlight the difficulty of predicting the BT component as only 2 of the 4 models were capable of yielding results higher than zero, contrasting with its ability to predict B and T components when they are issued independently, with a *f-1 weighted average score* of 0.7300 and 0.5836 respectively. The prototypical structure of the semantic wave of beginning at a T component, descending to a B, then rising again, is being modelled well, though rarer types of semantic wave are not.

6 Predicting Tutor Conversational Management Acts

The second decision process we model, given student SCMAs and an AC selected by a model such as that just described (though here using the gold standard AC tags), is the selection of one of the 12 either single or compound TCMAs observed in the corpus shown in table 7, or no TCMA (∅). This is a far more challenging task in terms of the sparsity of many of the classes, but we set up the results from our classifiers here as a starting point for future work. We again experiment with an HMM and CRF classifier.

6.1 HMM for Predicting TCMAs

The HMM classifier set-up for TCMA prediction is similar to that described in Section 5.1, only the observations are now drawn from the cross-product of the 13 possible SCMA type values and ACs, meaning there are 39 possible observation types when using the simple AC tags and 156 possible observation types when using the BEIS-enriched AC tags. In practice not all of these combinations were observed.

The underlying state model is a first-order MLE markov model of TCMA types plus ∅, resulting in 13 possible states excluding the start and end states. Viterbi decoding is used again at decoding time.

6.2 CRF Model for Predicting TCMAs

The CRF TCMA model uses similar input types to that described for AC prediction in Section 5.2, but instead of using just the SCMA tags, it also uses AC tags (either simple or BEIS). SCMA and AC inputs from previous timesteps were also used - we experimented with tags from both $t-1$ and $t-2$.

6.3 Results

The results for tagging TCMAs of our HMM and CRF models are presented in Table 10. The optimal model with the highest performance results in the three key performance measures is CRF-16, with

Model	Features	F1 macro	F1 weighted	Accuracy
Random	-	0.0337	0.1196	0.0777
HMM	AC, SCMA	0.0871	0.4002	0.5472
CRF-12	SCMA,sAC,Lex,Qm,Ul,Sw	0.1229	0.4199	0.5597
CRF-14	SCMA,sAC,Lex,Qm,Ul,Sw,Lem	0.1304	0.4197	0.5624
CRF-16	SCMA,ACbeis,Lex,Qm,Ul,Sw,Lem,Prev2	**0.1370**	**0.4239**	**0.5627**

Table 10: Performance for tagging Tutor Conversational Management Acts (TCMA) in a cross-validation. SCMA:Student Conversational Management Act, cAC:Compound Analogical Component, Lex:Lexical, Qm:Question Mark, Ul:Utterance Length, Sw:Stopwords Removal, Lem:Lemmatisation, cACbeis:Compound Analogical Component with BEIS, Prev2:2 previous simple SCMA + cAC steps

Model	w avg	NONE	QR	FIM	ABK
HMM AC, SCMA	0.4002	0.7096	0.4231	0.0000	0.0000
CRF-5 SCMA,cAC,Lex	0.4158	0.7153	0.6230	0.0167	0.1347
CRF-11 SMCA,sAC,Lex,Qm,Ul	0.4182	0.7153	0.6174	0.0300	0.1218
CRF-12 SCMA,sAC,Lex,Qm,Ul,Sw	0.4199	0.7162	0.7353	0.0300	0.1158
CRF-14 SCMA,sAC,Lex,Qm,Ul,Sw,Lem	0.4197	0.7158	0.7492	0.0300	0.1152
CRF-16 SCMA,ACbeis,Lex,Qm,Ul,Sw,Lem,Prev2	0.4239	0.7183	0.8052	0.0138	0.2443

Table 11: Breakdown Performance (F1-Score) for Tagging Tutor Conversational Management Acts (TCMAs) in a cross-validation. SCMA:Student Conversational Management Act, cAC:Compound Analogical Component, Lex:Lexical, Qm:Question Mark, Ul:Utterance Length, Sw:Stopwords Removal, Lem:Lemmatisation, cACbeis:Compound Analogical Component with BEIS, Prev2:2 previous simple SCMA + cAC steps

an *f-1 weighted average score* of 0.4239 an *f-1 macro average score* of 0.1370 and an *accuracy* of 0.5627. The features used by the optimal model are: SCMA, AC using substates (BIES), lexical, presence or absence of a question mark, student utterance length, removal of stopwords, lemmatisation and the inclusion of 2 previous SCMA and compound AC states (full states). This model easily surpasses the random baseline of *f-1 weighted average* of 0.1196. This random baseline demonstrates how the task of predicting TCMAs is a difficult challenge. The breakdown results in Table 11 show how the rare Question Response (QR) class and NONE, which represents the tutor issuing an utterance in their analogical explanation without a TCMA, yield *f-1 weighted average scores* of 0.8052 and 0.7183 respectively, demonstrating how this model could be implemented in an artificial tutoring agent if limited to these decisions. While ABK prediction accuracy rises slightly using the previous time-steps, the other TCMAs do not provide significant results, and other methods should be designed and tested.

7 Conclusions and Future Work

We have presented a novel annotation schema and sequence models for predicting analogical components (AC) and a tutor's CMAs in tutorial dialogue. As discussed, the models we have developed can be implemented in a dialogue system, theoretically providing a functional and coherent interactive experience to the student: the models are particularly effective in terms of the selection of the correct AC and the TCMAs of question response (QR) and NONE, representing the tutor issuing an utterance without any TCMA. The other TCMAs, while more frequent than QR, are not easily predictable with our current set-up and in future work will be modelled with alternative methods in such a way that minimal accuracy thresholds are achieved for use in a real tutorial system, including using a neural system with additional data or an information state update approach to dialogue management.

8 Acknowledgements

Del-Bosque-Trevino is partially supported by EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1) and CONACYT (National Council of Science and Technology of Mexico). Purver is partially supported by the EPSRC under grant EP/S033564/1, and by the European Union's Horizon 2020 programme under grant agreement 825153 (EMBEDDIA, Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this publication reflect only the authors' views and the Commission is not responsible for any use that may be made of the information it contains.

References

Vincent AWMM Aleven and Kenneth R Koedinger. 2002. An effective metacognitive strategy: learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, 26(2):147–179.

Mehrdad Alizadeh, Barbara Di Eugenio, Rachel Harsley, Nick Green, Davide Fossati, and Omar AlZoubi. 2015. A study of analogy in computer science tutorial dialogues. In *Proceedings of the 7th International Conference on Computer Supported Education*, volume 2, pages 232–237.

James Allen and Mark Core. 1997. Draft of damsl: Dialog act markup in several layers.

Kristy Elizabeth Boyer, Eun Young Ha, Michael D Wallis, Robert Phillips, Mladen A Vouk, and James C Lester. 2009. Discovering tutorial dialogue strategies with hidden markov models. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED)*, pages 141–148.

Kristy Elizabeth Boyer, Robert Phillips, Amy Ingram, Eun Young Ha, Michael Wallis, Mladen Vouk, and James Lester. 2011. Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden markov modeling approach. *International Journal of Artificial Intelligence in Education*, 21(1-2):65–81.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. In Lauren B Resnick, John M Levine, and Stephanie D Teasley, editors, *Perspectives on Socially Shared Cognition*, chapter 7, pages 127–149. American Psychological Association.

Paul Curzon, Peter McOwan, J Donohue, Seymour Wright, and William Marsh. 2018. Teaching computer science concepts. In Sue Sentance, Eric Barendsen, and Carsten Shulte, editors, *Computer Science Education: Perspectives on Teaching and Learning in School*, chapter 8, pages 91–108. Bloomsbury Publishing, London.

Jorge Del-Bosque-Trevino, Julian Hough, and Matthew Purver. 2020. Investigating the semantic wave in tutorial dialogues: An annotation scheme and corpus study on analogy components. In *Proceedings of the 24th SemDial Workshop on the Semantics and Pragmatics of Dialogue*, Virtually at Brandeis, Waltham, New Jersey. SEMDIAL.

Barbara Di Eugenio, Lin Chen, Nick Green, Davide Fossati, and Omar AlZoubi. 2013. Worked out examples in computer science tutoring. In *International Conference on Artificial Intelligence in Education*, pages 852–855. Springer.

Barbara Di Eugenio, Davide Fossati, Stellan Ohlsson, and David Cosejo. 2009. Towards explaining effective tutorial dialogues. In *Annual Meeting of the Cognitive Science Society*, pages 1430–1435.

Bruce Fraser. 1999. What are discourse markers? *Journal of Pragmatics*, 31(7):931–952.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Daniel Jurafsky and James H Martin. 2020. *Speech and Language Processing*, 3rd edition.

Karl Maton. 2013. Making semantic waves: A key to cumulative knowledge-building. *Linguistics and education*, 24(1):8–22.

Kaśka Porayska-Pomsta and Chris Mellish. 2013. Modelling human tutors' feedback to inform natural language interfaces for learning. *International Journal of Human-Computer Studies*, 71(6):703–724.

Vasile Rus, Nabin Maharjan, Lasang Jimba Tamang, Michael Yudelson, Susan R Berman, Stephen E Fancsali, and Steven Ritter. 2017. An analysis of human tutors' actions in tutorial dialogues. In *FLAIRS Conference*, pages 122–127.

Deborah Schiffrin. 1987. *Discourse markers*. 5. Cambridge University Press.

Lawrence Schourup. 1999. Discourse markers. *Lingua*, 107(3-4):227–265.

John R Searle. 1965. What is a speech act. *Perspectives in the philosophy of language: a concise anthology*, 2000:253–268.

Stephanie Ann Siler and Kurt VanLehn. 2009. Learning, interactional, and motivational outcomes in one-to-one synchronous computer-mediated versus face-to-face tutoring. *International Journal of Artificial Intelligence in Education*, 19(1):73–102.

Kurt VanLehn, Stephanie Siler, Charles Murray, Takashi Yamauchi, and William B Baggett. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3):209–249.