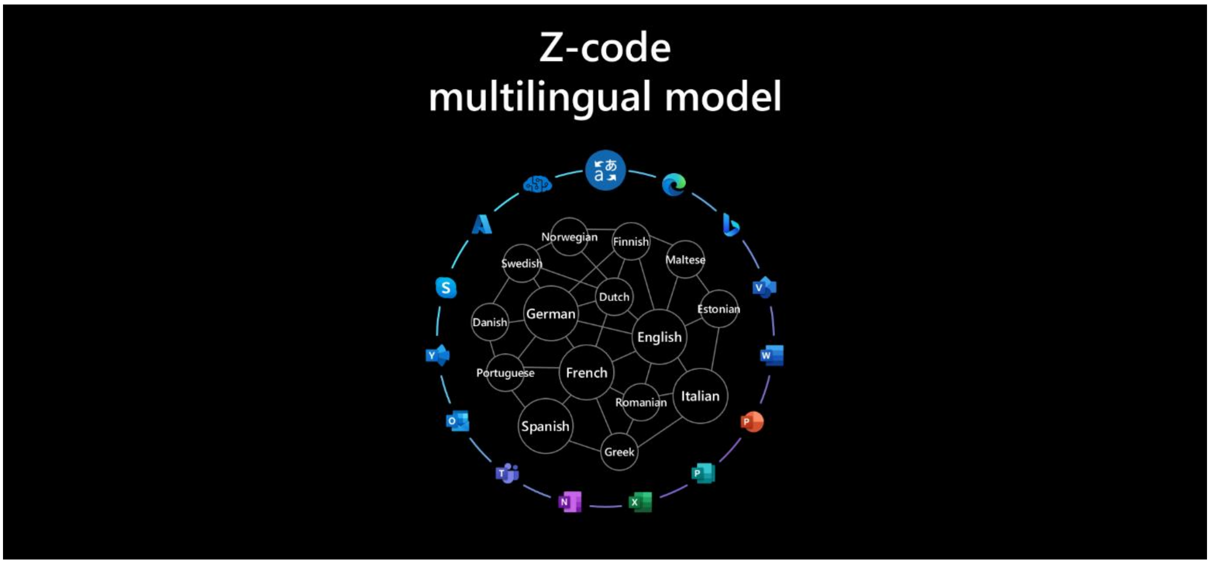
Multilingual translation at scale: 10000 language pairs and beyond

*Posted on **November 22, 2021**by **Microsoft Translator***



Microsoft is on a quest for **AI at Scale** with high ambition to enable the next generation of AI experiences. The Microsoft Translator **ZCode** team is working together with **Microsoft Project Turing** and Microsoft Research Asia to advance language and multilingual support at the core of this initiative. We continue to push frontiers with Multilingual models to support various language scenarios across Microsoft. Last summer, we announced our large scale **Multi-Lingual Mixture of Expert** model with **DeepSpeed** that can outperform individual large scale bi-lingual models. Recently, the latest Turing universal language representation model (**T-ULRv5**), a Microsoft-created model is once again the state of the art and at the top of the Google **XTREME public leaderboard** at that time. More recently, Microsoft announced the largest **Megatron-Turing NLG 530B** parameters model.

The annual Conference on Machine Translation (aka WMT 2021) concluded last week in beautiful Punta Cana, Dominican Republic. WMT brings together researchers from across the entire Machine Translation field, both industry and academia, to participate in a series of shared tasks, each defining a benchmark in an important area of machine translation to push the field into new frontiers.

The Microsoft Translator ZCode team, working together with Turing team and Microsoft Research Asia, competed in the "Large-scale Multilingual Translation" track, which consisted of a Full Task of translating between all 10,000 directions across 101 languages, and two Small tasks: One focused on 5 central and southern European languages, and one on 5 south-east Asian languages. The Microsoft ZCode-DeltaLM model won all three tasks by huge margins, including an incredible 10+ point gain over the M2M100 model in the large task evaluated on a massive 10,000 language pairs. (**Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation**, Wenzek et al, WMT 2021).

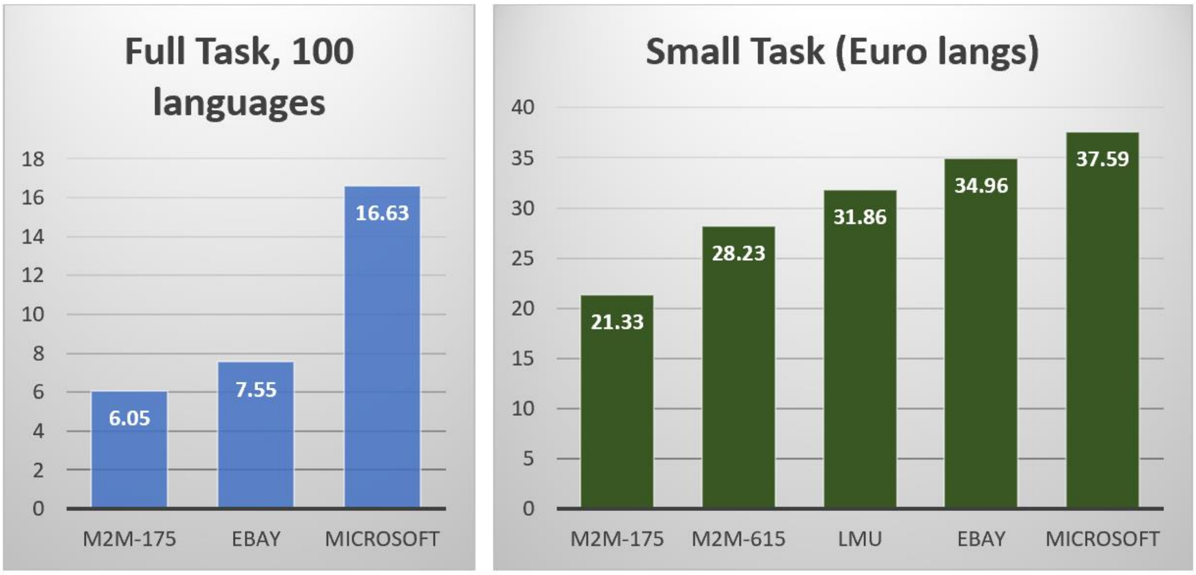
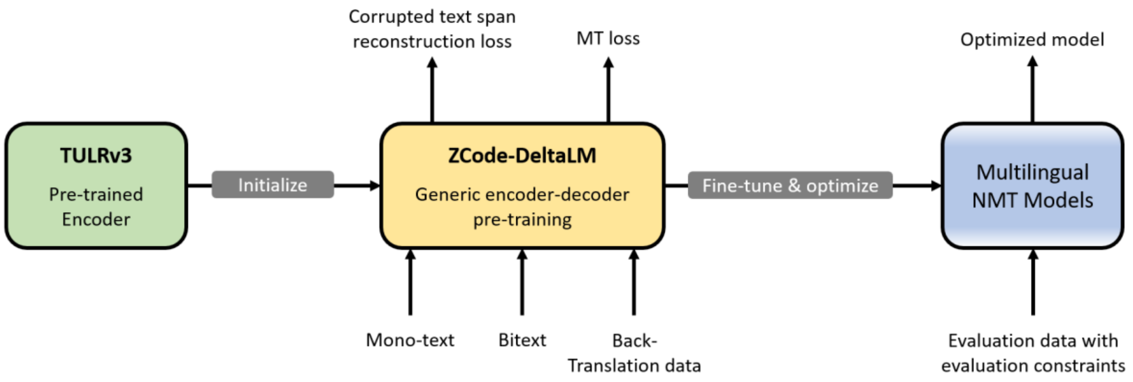


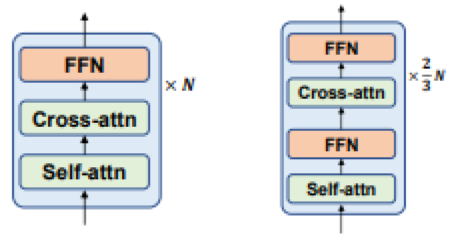
Figure 1: Official Results (BLEU scores) on the Full-Task and the Small-Task1 at the WMT 2021 Large Scale Multilingual Translation shared task

The ZCode-DeltaLM approach

In this blog post, let's take a look under the hood at the winning Microsoft ZCode-DeltaLM model. Our starting point was DeltaLM (**DeltaLM: Encoder-Decoder Pre-training for Language Generation and Translation by Augmenting Pretrained Multilingual Encoders**), the latest in the increasingly powerful series of massively multilingual pretrained language models from Microsoft.

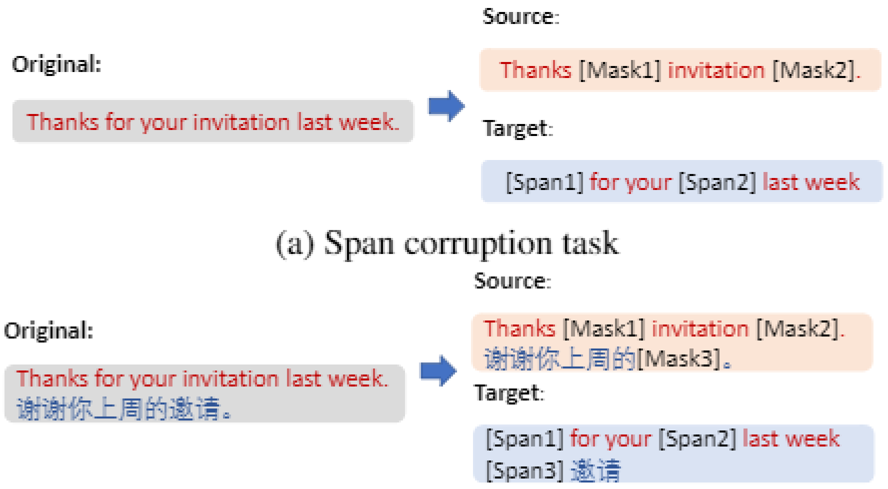


DeltaLM is an encoder-decoder model, but instead of training from scratch, it is initialized from a previously pretrained state-of-the-art encoder-only model, specifically (**TULRv3**). While initializing the encoder is straightforward, the decoder is less so, since it adds cross-attention to the encoder's self-attention. DeltaLM solves this problem with a novel interleaved architecture, where the self-attention and cross-attention alternate between layers, with the self-attention used in the odd layers and cross-attention used in the even layers. With this interleaving, the decoder structure matches the encoder, and so it can also be initialized the same way from TULRv3.



(a) Vanilla decoder (b) Interleaved decoder

DeltaLM is augmented by ZCode powerful multitask learning: **Multi-task Learning for Multilingual Neural Machine Translation**. Our models show that combining multitask and multilingual learning can significantly improve training for large scale pretrained language models. Such multitask multilingual learning paradigm is leveraging the inductive bias and regularization from several tasks and languages simultaneously to perform better on various downstream tasks. We are using translation task, denoising auto encoder task and translation span corruption task as shown in the figure below.


Original:
Thanks for your invitation last week.
Source:
Thanks [Mask1] invitation [Mask2].
Target:
[Span1] for your [Span2] last week
Original:
Thanks for your invitation last week.
谢谢你上周的邀请。
Source:
Thanks [Mask1] invitation [Mask2].
谢谢你上周的[Mask3]。
Target:
[Span1] for your [Span2] last week
[Span3] 邀请

(a) Span corruption task

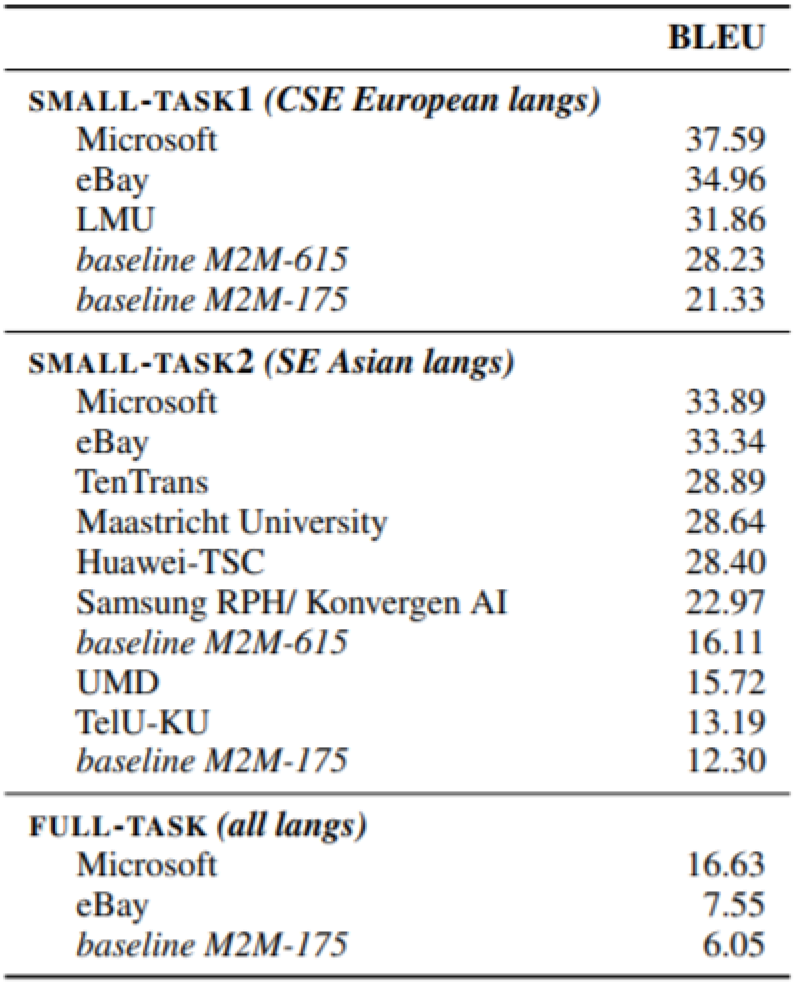
(b) Translation span corruption task

Winning the massively multilingual translation track

To build our winning massively multilingual translation system (**Multilingual Machine Translation Systems from Microsoft for WMT21 Shared Task**), we started with zCode-DeltaLM, and added a few tricks.

We apply progressive learning, first training a model with 24 encoder layers and 12 decoder layers, then continue training with 12 added encoder layers, resulting in a deep 36 layer encoder. To cover all language pairs, we generate dual-pseudo-parallel data where both sides of the parallel data are synthetic, translated by the model from English. We also apply iterative back-translation to generate synthetic data. We apply curriculum learning, starting with the entire noisy training data, then reducing it to a clean subset. We re-weight the translation objective to favor parallel data over the back-translation and dual-pseudo-parallel data. We apply temperature sampling to balance across language pairs. For each language pair, we choose, based on the dev set, whether to prefer direct translation or pivot translation through English.

Putting it all together, we knew we had an amazing massively multilingual system, but the official results on the blind test set exceeded our expectations. We scored 2.5 to 9 BLEU ahead of the next competitor, and 10 to 21 BLEU points ahead of the baseline M2M-175 model. On the dev test we compared against the larger M2M-615 model, which we also beat by 10 to 18 points.



	BLEU
SMALL-TASK1 *(CSE European langs)*	
Microsoft	37.59
eBay	34.96
LMU	31.86
baseline M2M-615	28.23
baseline M2M-175	21.33
SMALL-TASK2 *(SE Asian langs)*	
Microsoft	33.89
eBay	33.34
TenTrans	28.89
Maastricht University	28.64
Huawei-TSC	28.40
Samsung RPH/ Konvergen AI	22.97
baseline M2M-615	16.11
UMD	15.72
TelU-KU	13.19
baseline M2M-175	12.30
FULL-TASK *(all langs)*	
Microsoft	16.63
eBay	7.55
baseline M2M-175	6.05

Table 2: Official results for the three shared tasks in the large-scale multilingual machine translation task

Beyond Translation: Universal Language Generation

While we are excited about the big win at WMT 2021, what's even more exciting is that unlike the other competitors, our ZCode-DeltaLM model is not just a translation model, but rather a general pretrained encoder-decoder language model, usable for all kinds of generation tasks beyond translation. This really enable our models to perform quite well on various multilingual natural language generation tasks.

We reached a new SOTA in many popular generation tasks from **GEM Benchmark**, including Wikilingua (summarization), Text simplification (WikiAuto), and structure-to-text (WebNLG). The DeltaLM-ZCode model widely outperform much larger models such as mT5 XL (3.7B) which is also trained on much larger data as well. This demonstrated the efficiency and versatility of the models leading to strong performance across many tasks.

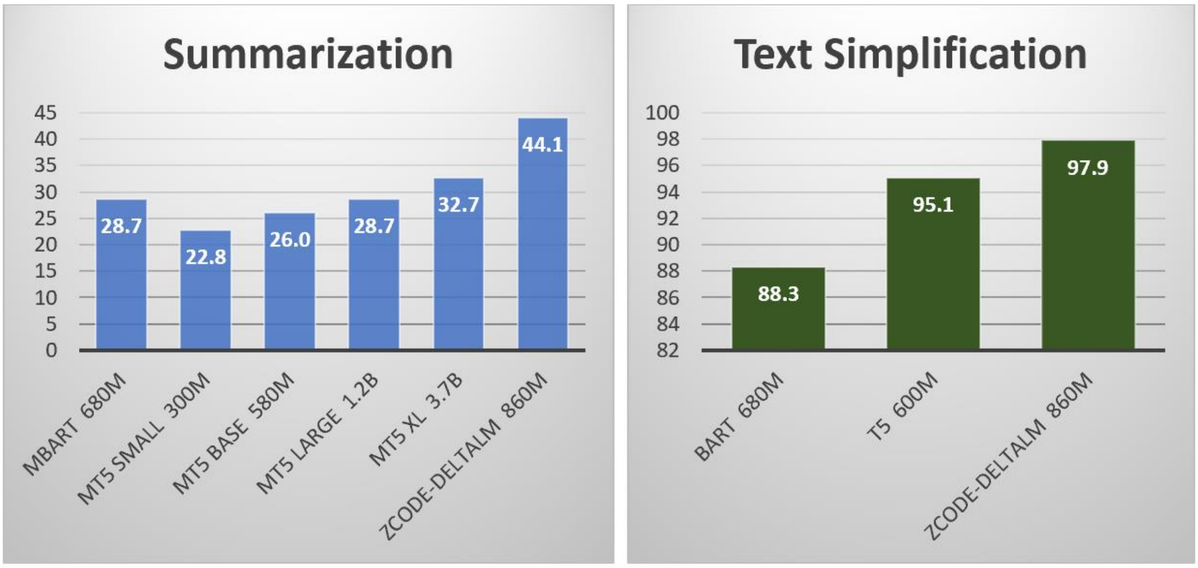


Figure 2. Performance (RL scores) of ZCode-DeltaLM on the Summarization and Text Simplification tasks in the GEM benchmark

Looking Ahead

Multilingual Machine Translation has reached a point where it performs very well, exceeding bilingual systems, on both low and high resource languages. Mixture of Experts (MoE) models have been shown to be a very good fit to scale up such models as has been shown in GShard. We explore how to efficiently scale such models with Mixture of Experts: **Scalable and Efficient MoE Training for Multitask Multilingual Models**. MoE models with massive multilingual data and unsupervised multitask training present unprecedent opportunity for such models to provide truly universal systems that can further enable the Microsoft Translator team to eliminate language barriers across the world, as well as support a variety of natural language generation tasks.

Acknowledgements

We would like to acknowledge and thank Francisco Guzman & his team who collected the massively multilingual FLORES test set and organized this WMT track with such large scale evaluation.

Translator Blog

Posted in **Research**

See more posts

BLOG POST

07 MAY 2021

Advancing sports analytics through AI research

SHARE

-
-
-

AUTHORS

- KT

  Karl Tuyls
- SO

  ShayeganOmidshafiei
- DH

  Daniel Hennes

- JC

  Jerome Connor
- ZW

  Zhe Wang
- ARC

  Adria RecasensContinente

Creating testing environments to help progress AI research out of the lab and into the real world is immensely challenging. Given AI's long association with games, it is perhaps no surprise that sports presents an exciting opportunity, offering researchers a testbed in which an AI-enabled system can assist humans in making complex, real-time decisions in a multiagent environment with dozens of dynamic, interacting individuals.

The rapid growth of sports data collection means we are in the midst of a remarkably important era for sports analytics. The availability of sports data is increasing in both quantity and granularity, transitioning from the days of aggregate high-level statistics and sabermetrics to more refined data such as event stream information (e.g., annotated passes or shots), high-fidelity player positional information, and on-body sensors. However, the field of sports analytics has only recently started to harness machine learning and AI for both understanding and advising human decision-makers in sports. In our recent paper published in collaboration with Liverpool Football Club (LFC) in JAIR, we envision the future landscape of sports analytics using a combination of statistical learning, video understanding, and game theory. We illustrate football, in particular, is a useful microcosm for studying AI research, offering benefits in the longer-term to decision-makers in sports in the form of an automated video-assistant coach (AVAC) system (Figure 1(A)).

Figure 1: (A) example illustration of an envisioned automated video-assistant coach interface, where attacking and defending players are detected, identified (in terms of player names), tracked, and subsequently passed into a predictive

trajectory model that can be used to analyse potential intents or prescribed trajectories. (B) stylised example of event detection, with a specific target event (e.g., kick) together with the deep learning model output (‘Signal’) evolving throughout the game.

Football - an interesting opportunity for AI

In comparison to some other sports, football has been rather late with starting to systematically collect large sets of data for scientific analytics purposes aiming to progress teams’ gameplay. This is for several reasons, with the most prominent being that there are far less controllable settings of the game compared to other sports (large outdoor pitch, dynamic game, etc.), and also the dominant credo to rely mainly on human specialists with track records and experience in professional football. On these lines, Arrigo Sacchi, a successful Italian football coach and manager who never played professional football in his career, responded to criticism over his lack of experience with his famous quote when becoming a coach at Milan in 1987: “I never realised that to be a jockey you had to be a horse first.”

Football Analytics poses challenges that are well suited for a wide variety of AI techniques, coming from the intersection of 3 fields: computer vision, statistical learning and game theory (visualised in Figure 2). While these fields are individually useful for football analytics, their benefits become especially tangible when combined: players need to take sequential decision-making in the presence of other players (cooperative and adversarial) and as such game theory, a theory of interactive decision making, becomes highly relevant. Moreover, tactical solutions to particular in-game situations can be learnt based on in-game and specific player representations, which makes statistical learning a highly relevant area. Finally, players can be tracked and game scenarios can be recognised automatically from widely-available image and video inputs.

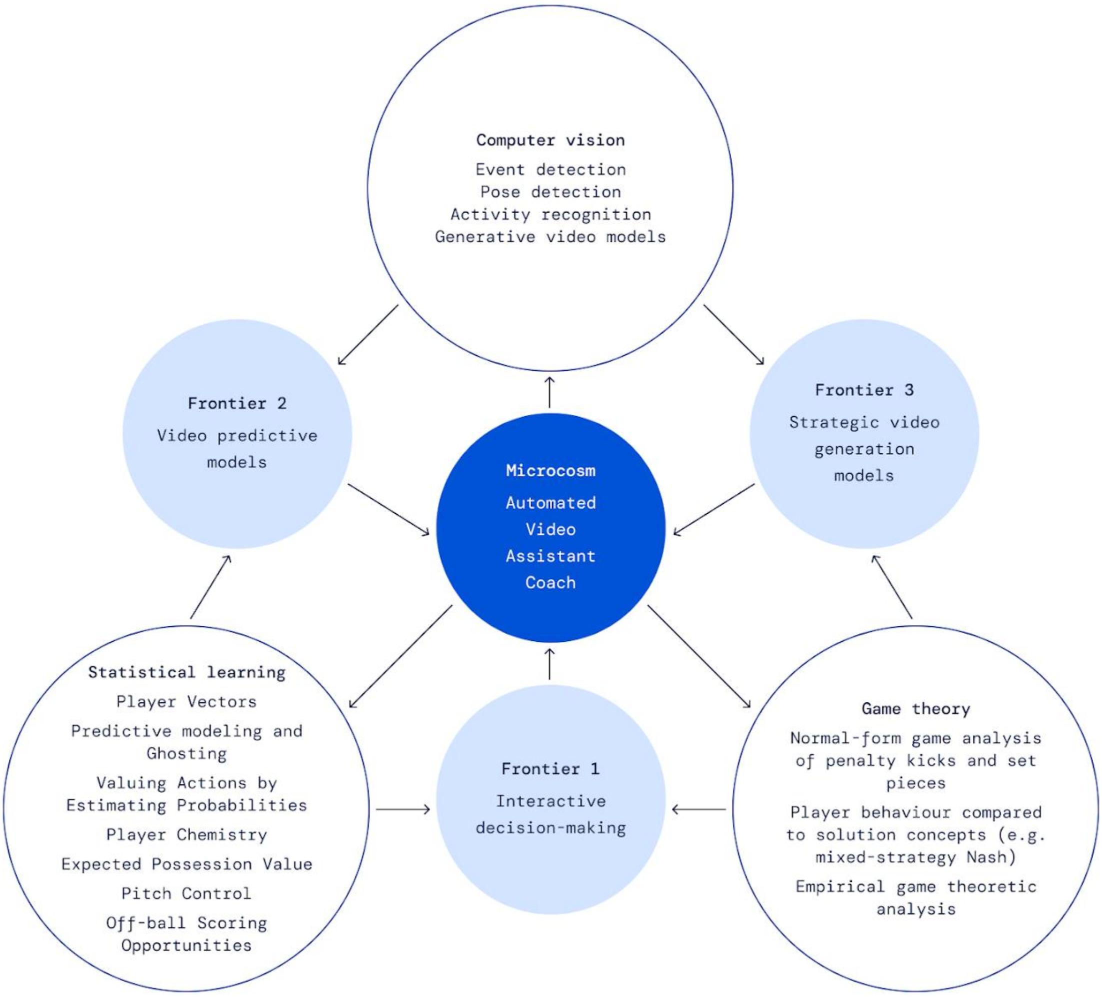

Computer vision
Event detection
Pose detection
Activity recognition
Generative video models
Frontier 2
Video predictive models
Microcosm
Automated Video Assistant Coach
Frontier 3
Strategic video generation models
Statistical learning
Player Vectors
Predictive modeling and Ghosting
Valuing Actions by Estimating Probabilities
Player Chemistry
Expected Possession Value
Pitch Control
Off-ball Scoring Opportunities
Frontier 1
Interactive decision-making
Game theory
Normal-form game analysis of penalty kicks and set pieces
Player behaviour compared to solution concepts (e.g. mixed-strategy Nash)
Empirical game theoretic analysis

Figure 2: illustrative overview of the three key fields (Game Theory, Statistical Learning, and Computer Vision) that have played an important role in advancing the state of football analytics (with examples from literature listed in each associated domain, and associated overlapping frontiers indicated).

The AVAC system we envision is situated within the microcosm that is formed by the intersection of these three research fields (Figure 2). In our research in this exciting domain, we not only lay out a roadmap for scientific and engineering problems that can be tackled for years to come, but we also present new original results at the crossroads of game theoretic analysis, statistical learning, and computer vision to illustrate what this exciting area has to offer to football.

How AI could help football

Game theory plays an important role in the study of sports, enabling theoretical grounding of players' behavioral strategies. In the case of football, many of its scenarios can actually be modeled as zero-sum games, which have been studied extensively since the inception of game theory. For example, here we model the penalty kick situation as a two-player asymmetric game, where the kicker's strategies may be neatly categorised as left, center, or right shots. To study this problem, we augment game-theoretic analysis in the penalty kick scenario with Player Vectors, which summarise the playing styles of individual football players. With such representations of individual players, we are able to group kickers with similar playing styles, and then conduct game-theoretic analysis on the group-level (Figure 3). Our results show that the identified shooting strategies of different groups are statistically distinct. For example, we find that one group prefers to shoot to the left corner of the goal mouth, while another tends to shoot to the left and right corners more evenly. Such insights may help goalkeepers diversify their defense strategies when playing against different types of players. Building on this game-theoretic view, one can consider the durative nature of football by analysing it in the form of temporally-extended games, use this to advise tactics to individual players, or even go further to optimise the overall team strategy.

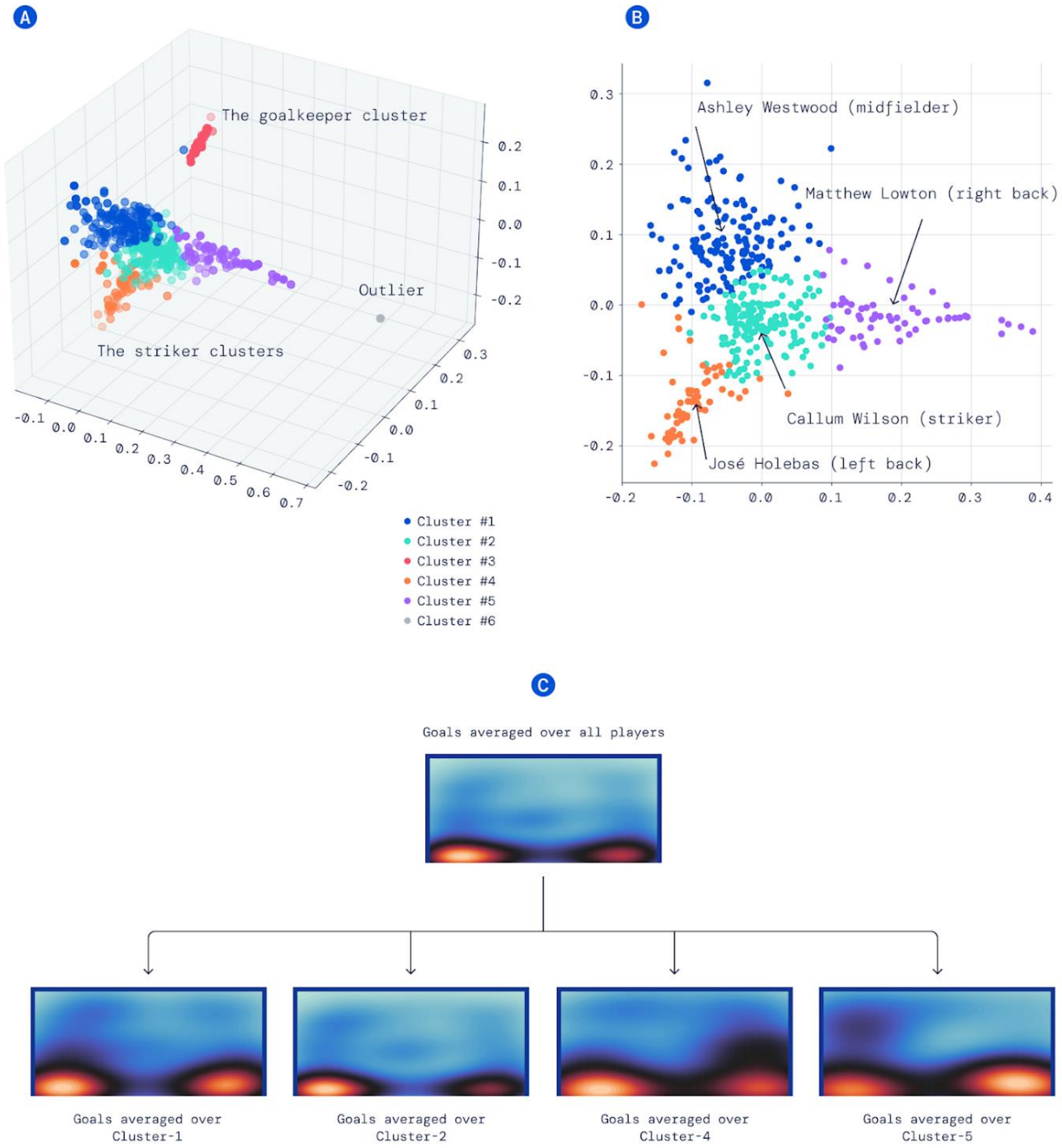


Figure 3: (A) and (B) visualise clusters of Player Vectors, for players in an example database of over 12000 penalty kicks. Using such a characterisation of player behaviours, one can visualise associated heatmaps of goals by kickers in various clusters, as illustrated in (C).

On the side of statistical learning, representation learning has yet to be fully exploited in sports analytics, which would enable informative summarisation of the behavior of individual players and football teams. Moreover, we believe that the interaction between game theory and statistical learning would catalyse advances in sports analytics further. In the above penalty kick scenario, for instance, augmenting the analysis with player-specific statistics (Player Vectors) provided deeper insights into how various types of players behave or make decisions about their actions in the penalty kick scenario. As another example of this, one can study 'ghosting', which refers to a particular data driven analysis of how players should have acted in hindsight in sports analytics (which bears connections to the notion of regret in online learning and game theory). The ghosting model suggests alternative player trajectories for a given play, e.g., based on the league average or a selected team. Predicted trajectories are usually visualised as a translucent layer over the original play, hence the term 'ghosting' (see Figure 4 for a visual example). Generative trajectory prediction models allow us to gain insights by analysing key situations of a game and how they might have played out differently. These models also bear potential in predicting the implications of a tactical change, a key player's injury, or substitution on the own team's performance along with the opposition's response to such a change.

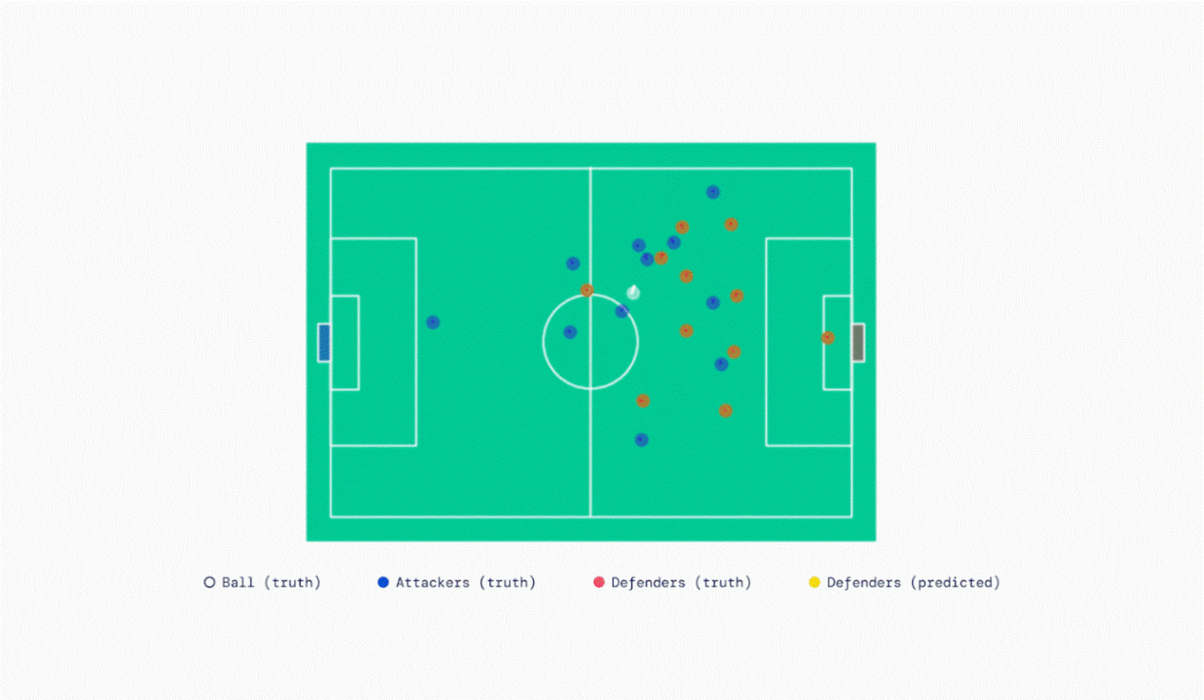


Figure 4: Example of predictive modelling using football tracking data. Here, the ground truth data for the ball, attackers, and defenders is visualised in addition to defender predictions made by a sequential-predictive trajectory model.

Finally, we consider computer vision to be one of the most promising avenues for advancing the boundaries of state of the art sports analytics research. By detecting events purely from video, a topic that has been well-studied in the computer vision community (e.g., see the following survey and our paper for additional references), the potential range of application is enormous. By associating events with particular frames, videos become searchable and ever more useful (e.g., automatic highlight generation becomes possible). Football video, in turn, offers an interesting application domain for computer vision. The large numbers of football videos satisfies a prerequisite for modern AI techniques. While each football video is different, the settings do not vary greatly, which makes the task ideal for sharpening AI algorithms. Third-party providers also exist to furnish hand-labelled event data that can be useful in training video models and are time consuming to generate, so both supervised and unsupervised algorithms can be used for football event detection. Figure 1(B), for example, provides a stylised visualisation of a deep learning model trained with supervised methods to recognise target events (e.g., kicks) purely from video.

The application of advanced AI techniques to football has the potential to revolutionise the game across many axes, for players, decision-makers, fans, and broadcasters. Such advances will also be important as they also bear potential to further democratise the sport itself (e.g., rather than relying on judgement calls from in-person scouts/experts, one may use techniques such as computer vision to quantify skillsets of players from under-represented regions, those from lower-level leagues, etc.). We believe that the development of increasingly advanced AI techniques afforded by the football microcosm might be applicable to broader domains. To this end, we are co-organising (with several external organisers) an IJCAI 2021 workshop on AI for Sports Analytics later this year, which we welcome interested researchers to attend. For researchers interested in this topic, publicly available datasets have been made available both by analytics companies such as StatsBomb (dataset link) and the wider research community (dataset link). Furthermore, the paper provides a comprehensive overview of research in this domain.

Paper and related links:

- JAIR Paper
- IJCAI 2021 AI for Sports Analytics virtual workshop

Work done as a collaboration with contributors: Karl Tuyls, ShayeganOmidshafiei, Paul Muller, Zhe Wang, Jerome Connor, Daniel Hennes, Ian Graham, William Spearman, Tim Waskett, Dafydd Steele, Pauline Luc, Adria Recasens, Alexandre Galashov, Gregory Thornton, Romuald Elie, Pablo Sprechmann, Pol Moreno, Kris Cao, Marta Garnelo, Praneet Dutta, Michal Valko, Nicolas Heess, Alex Bridgland, Julien Perolat, Bart De Vylder, Ali Eslami, Mark Rowland, Andrew Jaegle, Yi Yang, Remi Munos, Trevor Back, Razia Ahamed, Simon Bouton, Nathalie Beauguerlange, Jackson Broshear, Thore Graepel, and Demis Hassabis.

Further reading

PUBLICATION

Game Plan: What AI can do for Football, and What Football can do for AI

Karl Tuyls, ShayeganOmidshafiei, et al. The Journal of Artificial Intelligence Research 2020

DOWNLOAD

PUBLICATION

Navigating the Landscape of Games

ShayeganOmidshafiei, Karl Tuyls, et al. arXiv 2020

DOWNLOAD